

**IN THE UNITED STATES DISTRICT COURT  
FOR THE WESTERN DISTRICT OF TEXAS  
WACO DIVISION**

**XOCKETS, INC.,**

Plaintiff,

v.

**NVIDIA CORPORATION,  
MICROSOFT CORPORATION, and  
RPX CORPORATION,**

Defendants.

Civil Action No. 6:24-cv-453

**JURY TRIAL DEMANDED**

**FIRST AMENDED COMPLAINT FOR VIOLATION OF FEDERAL ANTITRUST  
LAWS AND PATENT INFRINGEMENT, AND REQUEST FOR INJUNCTION**

## TABLE OF CONTENTS

<b>INTRODUCTION.....</b>	<b>4</b>
I.    Xockets and Its DPU Inventions.....	4
II.   Xockets Presented Its Patented DPU Technology, Which Was Then Stolen By Defendants .....	8
III.  Xockets’ DPU Inventions Are Essential Elements of NVIDIA’s and Microsoft’s GPU-Enabled Server Computer Systems.....	12
IV.   NVIDIA’s and Microsoft’s Predatory Infringement Threatens Xockets and Its Exclusive Patent Rights.....	20
<b>THE PARTIES.....</b>	<b>22</b>
I.    Xockets .....	22
II.   NVIDIA .....	24
III.  Microsoft.....	24
IV.   RPX.....	24
<b>JURISDICTION AND VENUE.....</b>	<b>24</b>
I.    NVIDIA .....	25
II.   Microsoft.....	27
III.  RPX.....	29
<b>JOINDER.....</b>	<b>30</b>
<b>FACTUAL ALLEGATIONS.....</b>	<b>31</b>
I.    Xockets Patents .....	31
A.    The New Cloud Processor Patents .....	31
(1)    The ’209 Patent – DPU Computing Architecture, Security .....	31
(2)    The ’924 Patent – DPU Network Overlay, Security .....	35
(3)    The ’350 Patent – DPU Stream Processing .....	38
B.    The New Cloud Fabric Patents .....	42
(1)    The ’297 Patent – DPU Cloud Network Fabric .....	43
(2)    The ’161 Patent – DPU Cloud Network Fabric .....	46
(3)    The ’092 Patent – DPU In-Network Computing .....	48
(4)    The ’640 Patent – DPU In-Network Computing .....	52
II.   Background on Xockets’ Cloud Computing Inventions .....	55
III.  NVIDIA’s Use of Xockets’ Patented Technology.....	63

A.	NVIDIA’s Infringement of the New Cloud Processor Patents .....	64
B.	NVIDIA’s Infringement of the New Cloud Fabric Patents .....	72
C.	NVIDIA’s Knowledge of the XOCKETS Patents.....	92
IV.	Microsoft’s Use of Xockets’ Technology .....	93
A.	Microsoft’s Praise of Xockets’ Patented technology.....	95
B.	Microsoft’s Knowledge of the XOCKETS Patents .....	97
V.	Representative Benefits of Xockets’ Patented Technology.....	100
VI.	RPX’s Business.....	104
VII.	NVIDIA and Microsoft Respond to Xockets’ 2024 Fundraising Efforts By Creating a Buyers’ Cartel.....	106
VIII.	Illegal Agreement Between the Defendants .....	108
<b>COUNT I: VIOLATION OF SECTION 1 OF THE SHERMAN ACT BASED ON DEFENDANTS’ CONSPIRACY IN RESTRAINT OF TRADE .....</b>		<b>110</b>
<b>COUNT II: VIOLATION OF SECTION 2 OF THE SHERMAN ACT BASED ON DEFENDANTS’ CONSPIRACY TO CREATE OR MAINTAIN A MONOPSONY.....</b>		<b>111</b>
<b>COUNT III: INFRINGEMENT OF THE ’209 PATENT .....</b>		<b>113</b>
I.	Direct Infringement.....	113
A.	NVIDIA’s Direct Infringement.....	114
B.	Microsoft’s Direct Infringement .....	116
II.	Indirect Infringement .....	117
A.	NVIDIA’s Indirect Infringement .....	117
B.	Microsoft’s Indirect Infringement.....	120
III.	Willful Infringement .....	122
A.	NVIDIA’s Willful Infringement .....	122
B.	Microsoft’s Willful Infringement .....	123
<b>COUNT IV: INFRINGEMENT OF THE ’924 PATENT .....</b>		<b>124</b>
I.	Direct Infringement.....	124
A.	NVIDIA’s Direct Infringement.....	125
B.	Microsoft’s Direct Infringement .....	126
II.	Indirect Infringement .....	128
A.	NVIDIA’s Indirect Infringement .....	128
B.	Microsoft’s Indirect Infringement.....	130
III.	Willful Infringement .....	133

A.	NVIDIA’s Willful Infringement .....	133
B.	Microsoft’s Willful Infringement .....	134
<b>COUNT V: INFRINGEMENT OF THE ’350 PATENT .....</b>		<b>134</b>
I.	Direct Infringement.....	135
A.	NVIDIA’s Direct Infringement.....	136
B.	Microsoft’s Direct Infringement .....	137
II.	Indirect Infringement .....	139
A.	NVIDIA’s Indirect Infringement .....	139
B.	Microsoft’s Indirect Infringement.....	141
III.	Willful Infringement .....	144
A.	NVIDIA’s Willful Infringement .....	144
B.	Microsoft’s Willful Infringement .....	145
<b>COUNT VI: INFRINGEMENT OF THE ’297 PATENT .....</b>		<b>145</b>
I.	Direct Infringement.....	146
A.	NVIDIA’s Direct Infringement.....	147
B.	Microsoft’s Direct Infringement .....	148
II.	Indirect Infringement .....	149
A.	NVIDIA’s Indirect Infringement .....	149
B.	Microsoft’s Indirect Infringement.....	151
III.	Willful Infringement .....	154
A.	NVIDIA’s Willful Infringement .....	154
B.	Microsoft’s Willful Infringement .....	155
<b>COUNT VII: INFRINGEMENT OF THE ’161 PATENT .....</b>		<b>155</b>
I.	Direct Infringement.....	156
A.	NVIDIA’s Direct Infringement.....	156
B.	Microsoft’s Direct Infringement .....	158
II.	Indirect Infringement .....	159
A.	NVIDIA’s Indirect Infringement .....	159
B.	Microsoft’s Indirect Infringement.....	161
III.	Willful Infringement .....	164
A.	NVIDIA’s Willful Infringement .....	164
B.	Microsoft’s Willful Infringement .....	165
<b>COUNT VIII: INFRINGEMENT OF THE ’092 PATENT .....</b>		<b>165</b>

I.	Direct Infringement.....	166
A.	NVIDIA’s Direct Infringement.....	167
B.	Microsoft’s Direct Infringement .....	168
II.	Indirect Infringement .....	169
A.	NVIDIA’s Indirect Infringement .....	169
B.	Microsoft’s Indirect Infringement.....	172
III.	Willful Infringement .....	174
A.	NVIDIA’s Willful Infringement .....	174
B.	Microsoft’s Willful Infringement .....	175
<b>COUNT IX: INFRINGEMENT OF THE ’640 PATENT .....</b>		<b>176</b>
I.	Direct Infringement.....	176
A.	NVIDIA’s Direct Infringement.....	177
B.	Microsoft’s Direct Infringement .....	178
II.	Indirect Infringement .....	179
A.	NVIDIA’s Indirect Infringement .....	179
B.	Microsoft’s Indirect Infringement.....	182
III.	Willful Infringement .....	184
A.	NVIDIA’s Willful Infringement .....	184
B.	Microsoft’s Willful Infringement .....	185
<b>INJUNCTIVE RELIEF .....</b>		<b>186</b>
<b>DAMAGES .....</b>		<b>193</b>
<b>ATTORNEYS FEES.....</b>		<b>193</b>
<b>DEMAND FOR JURY TRIAL.....</b>		<b>194</b>
<b>PRAYER FOR RELIEF.....</b>		<b>194</b>

Plaintiff Xockets, Inc. (“Plaintiff” or “Xockets”) hereby submits this First Amended Complaint against Defendants NVIDIA Corporation (“NVIDIA”), Microsoft Corporation (“Microsoft”), and RPX Corporation (“RPX”) for violation of federal antitrust laws and patent infringement for which it seeks an injunction, and states as follows:

1. NVIDIA holds monopoly power in the market for GPU-enabled artificial intelligence computer systems, holding market share above 90% by unit. Microsoft has combined its leading position in cloud services with control over leading generative AI models to maintain and/or create a monopoly in GPU-enabled generative artificial intelligence platforms. Microsoft has entered into unlawful agreements in restraint of trade with the owner of leading generative AI models, including OpenAI, and with the dominant supply of GPU-enabled artificial intelligence computer systems. NVIDIA and Microsoft have formed a cartel to create and/or maintain a monopoly in GPU-enabled generative artificial intelligence. As part of this cartel, NVIDIA and Microsoft have formed a buyers’ cartel to avoid paying the fair market price for the fundamental intellectual property that transformed GPU and GPU-enabled platforms from a niche product for gamers and cryptocurrency miners into the most important industrial component in the United States economy today. The fundamental technology was created by Xockets. This buyers’ cartel is designed to fix below market level the price for the critical technology held by Xockets.

2. The buyers’ cartel at issue in this case is part of a pattern of illegal cartel behavior engaged in by NVIDIA and Microsoft, as evidenced by the ongoing investigations of these entities by the United States Department of Justice, the United States Federal Trade Commission, and the European Union.

3. The illegal buyers’ cartel in this case was facilitated by RPX. RPX was formed at the request of Big Tech companies to enable and create buyers’ cartels for intellectual property.

RPX previously touted on its website that “[i]n effect, RPX can buy ‘wholesale’ on behalf of our client network, while our clients otherwise would pay ‘retail’ if transacting on their own.” In subsequent years RPX tried to erase this admission from the web because the “client network” as RPX euphemistically describes it consists of the companies that hold monopoly power in essentially every high technology field in this Country. This includes NVIDIA and Microsoft, who engaged RPX to manage the buyers’ cartel for the fundamental intellectual property that drives the AI revolution.

4. The fundamental intellectual property that turned NVIDIA’s GPUs from niche equipment for gamers and cryptocurrency miners to the driver of the AI revolution was created by Xockets. Every generation NVIDIA makes incremental improvements in its GPUs (including its Hopper and upcoming Blackwell GPUs) through a combination of leveraging improvements in manufacturing technology of the companies which manufacture its chips and minor improvements in its architecture.

5. What has allowed NVIDIA to monopolize the field of GPU-enabled artificial intelligence servers that third parties can buy is the introduction of three other entirely different components that use Xockets’ patented technology. These components, or Data Processing Units (DPUs), include the BlueField, ConnectX, and NVLink Switch DPUs for offloading, accelerating, and isolating data-intensive workloads from server processors in cloud data centers. They are necessary for allowing NVIDIA to combine a large number of GPU-enabled server boards or modules in order to process the vast amounts of data necessary for artificial intelligence and to provide this technology to customers across the web. NVIDIA’s CEO described the importance of its DPUs:

“The holy trinity of computing ... is the CPU, the GPU, and the DPU. These three processors are fundamental to computing.”<sup>1</sup>

“The data center has become the new unit of computing. DPUs are an essential element of modern and secure accelerated data centers in which CPUs, GPUs and DPUs are able to combine into a single computing unit that’s fully programmable, AI-enabled and can deliver levels of security and compute power not previously possible.”<sup>2</sup>

“A single BlueField-2 DPU can deliver the same data center services that could consume up to 125 CPU cores.”<sup>3</sup>

6. NVIDIA’s CEO also described the critical importance of its NVLink Switch DPUs: “for large language models like the Chat GPT and others like it . . . all these GPUs have to share the results, partial products [of AI model training operations] . . . Whenever they do all-to-all, all-gather, whenever they communicate with each other, that NVLink Switch is communicating almost 10 times faster than what we could do in the past using the fastest networks.”<sup>4</sup> “The miracle is this chip—this NVLink Chip.”<sup>5</sup>

7. Since the introduction of Xockets’ patented designs for its DPUs, NVIDIA’s market capitalization has exploded, from \$180 billion to approximately \$3 trillion.

8. NVIDIA did not invent the technology in its BlueField, ConnectX, and NVLink Switch DPUs. This technology was taken from Xockets. And it was done so knowingly. Instead

---

<sup>1</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

<sup>2</sup> <https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center>.

<sup>3</sup> <https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center>.

<sup>4</sup> GTC March 2024 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=Y2F8yisiS6E&t=3403s> (56:43–59:06).

<sup>5</sup> NVIDIA CEO Jensen Huang Keynote at COMPUTEX 2024, <https://www.youtube.com/watch?v=pKXDVsWZmUU&t=4283s> (1:11:23–1:15:57).



of paying fair value for the technology, Microsoft and NVIDIA took it without permission. And when NVIDIA and Microsoft were put on notice that Xockets would not allow its technology to be used illegally without a license they formed an illegal buyers' cartel with RPX to drive the price of Xockets' patents on its DPU inventions below the market price and/or drive Xockets out of business. This lawsuit will stop the illegal conduct.

9. This lawsuit will also stop the unlawful use of Xockets' intellectual property. Xockets seeks to enjoin the sale of all Accused Products<sup>6</sup> as a remedy for their willful and unlawful behavior, including enjoining the release of NVIDIA's new Blackwell GPU-enabled server computer systems and Microsoft's use thereof for generative AI platforms.

## **INTRODUCTION**

### **I. XOCKETS AND ITS DPU INVENTIONS**

10. Xockets was founded in 2012 by Dr. Parin Dalal and a team of network infrastructure engineers, turned early cloud engineers. Dr. Dalal received a bachelor's degree in computer science from University of California, Berkeley and his Ph.D. in theoretical physics from the University of California, San Diego, and began his career as an engineer designing CPUs and GPUs. Today, Dr. Dalal is Principal Engineer, Machine Learning and Artificial Intelligence, at Google. Prior to joining Google Dr. Dalal led company-wide strategic AI decision-making at Varian Medical Systems, now a Siemens company, as its Vice President of Advanced AI developing AI-based formulations of cancer treatments to save lives.

11. Founding investors in Xockets include the current CTO of Intel, Dr. Greg Lavender, who was a long-time Xockets Board member and considered Xockets' DPU architecture a

---

<sup>6</sup> See paragraphs 84–169 for Patents, and paragraphs 188, 209, and 244 for Accused Products.

“transformational” and “revolutionary” new computing architecture in clouds; Robert Cote, one of the nation’s leading IP investors and lawyers, who guided the company in ensuring that Xockets’ breakthrough DPU inventions were protected by United States patents; Jerry Yang, cofounder of Yahoo, who invested through AME cloud ventures, a venture capital firm he founded to invest in breakthrough cloud technologies.

12. In the early 2010s, Xockets’ co-founder and lead inventor, Dr. Parin Dalal, had the vision to see that conventional wisdom in the computing industry—that relies on expected increases in transistor density for ever-faster computing performance known as Moore’s Law—would fail to meet the unique challenges that data-intensive workloads would pose to distributed computing performance in cloud data centers. He foresaw that Moore’s Law would end as the data workloads driving distributed computing in clouds would grow by orders of magnitude. The manipulation and analysis of vast data sets across GPU-enabled server processors in the training of large language models, like ChatGPT, requires that this problem be addressed to enable the age of artificial intelligence that is now underway in the world.

13. Dr. Dalal realized that a new computing paradigm, involving a new accelerated computing architecture that extends into the network of cloud data centers, was needed. This new computing architecture for processing data-intensive workloads was implemented by Xockets in a new cloud processor known today as a Data Processing Unit, or DPU. It is designed to provide flexible hardware-like handling of computing operations at the speed of the network—or line rate—with software-like programmability that can form programmable logic pipelines of hardware accelerators for processing data-intensive workloads independent of server processors and conventional computing architectures. This programmable hardware acceleration in the network invented by Dr. Dalal can run new, varied, and evolving cloud infrastructure services. It provides

the versatility clouds require to offload infrastructure services and accelerate many different kinds of data-intensive workloads and processes, freeing up server processors to run their main workloads or applications for customers at ever-increasing speeds and lower power costs.

14. Xockets described Dr. Dalal's inventions in a series of patent applications filed with the United States Patent Office beginning in May 2012, and did so in reliance on the promise made in the United States Constitution that Xockets would be granted exclusive rights to Dr. Dalal's DPU inventions. These exclusive rights were placed by Congress in the United States Patent Laws and are fundamental to the innovation economy that our nation's Founders sought to build in a country that was once a startup nation. With this promise, people from all over the world and from all walks of life came to this country and created an innovation economy that is unparalleled in history—from which was built the largest economy and most prosperous nation on earth.

15. The growth of this innovation economy depends on the strict enforcement of an inventor's exclusive rights as promised in the United States Constitution. Indeed, the most important inventions in our nation's history have come from entrepreneurs like Dr. Dalal and startups like Xockets, not from those who hold the reins of power. The strict enforcement of intellectual property rights is what creates a level playing field for inventors and incentivizes the personal sacrifice that these entrepreneurs must make to build a future to benefit us all, and it is what instills confidence in investors to invest in breakthrough new startups like Xockets and inventors like Dr. Dalal.

16. To date, Xockets has obtained a number of patents covering many aspects of Dr. Dalal's DPU inventions—as there were many problems to solve. Xockets currently has over 60 patent applications prepared for filing directed to numerous other DPU inventions. Xockets' issued patents include: (i) Xockets' DPU Computing Architecture Patents (also known as the “*New Cloud*

*Processor Patents*”), including U.S. Patent Nos. 11,080,209 (“the ’209 Patent” – DPU Computing Architecture, Security), U.S. Patent No. 10,649,924 (“the ’924 Patent” – DPU Network Overlay, Security), and U.S. Patent No. 11,082,350 (“the ’350 Patent” – DPU Stream Processing); and (ii) Xockets’ DPU Switching Architecture Patents (also known as the “*New Cloud Fabric Patents*”), including U.S. Patent No. 10,223,297 (“the ’297 Patent” – DPU Cloud Network Fabric), U.S. Patent No. 9,378,161 (“the ’161 Patent” – DPU Cloud Network Fabric), U.S. Patent No. 10,212,092 (“the ’092 Patent” – DPU In-Network Computing), and U.S. Patent No. 9,436,640 (“the ’640 Patent” – DPU In-Network Computing). These patents, including the ’209 Patent, ’924 Patent, ’350 Patent, ’297 Patent, ’161 Patent, ’092 Patent, and ’640 Patent, are collectively referred to herein as the “Asserted Patents” or “Xockets Patents.” In addition to being infringed, the Xockets Patents are targets of Defendants’ unlawful buyers’ cartel.

17. Xockets’ patented inventions include a groundbreaking new cloud computing architecture and a new cloud network fabric—a reinvention of cloud distributed computing from the ground up—that serve to dramatically increase the speed and lower the costs of distributed computing services and AI. To do so, Xockets’ patented DPU architecture enables cloud server computers to offload, accelerate, and isolate critical data-intensive tasks that would otherwise overburden server processors.

18. Xockets’ New Cloud Processor Patents, for example, describe a virtual switch computing architecture for offloading from server processors to DPUs, or offload processor modules, accelerating, and isolating data-intensive workloads in clouds such as security, networking, and storage computing operations in moving data between server processors. In other words, Xockets’ Patents describe the use of a virtual switch computing architecture in a new cloud processor for brokering collective communication between server processors (the movement of

data between CPUs, GPUs, and hybrids of these server processors) independent of the limitations of conventional computing architectures, and for freeing up server processors to run customer applications at higher speeds and lower cost.

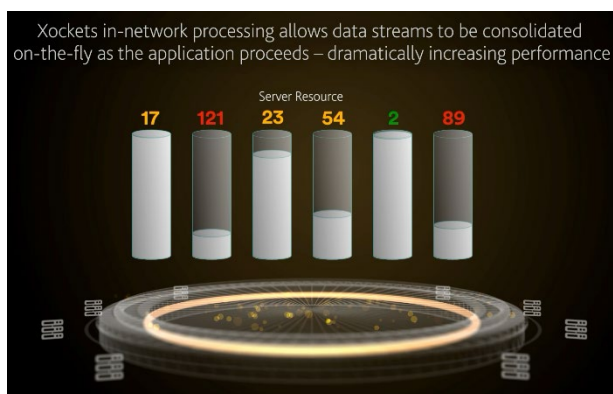
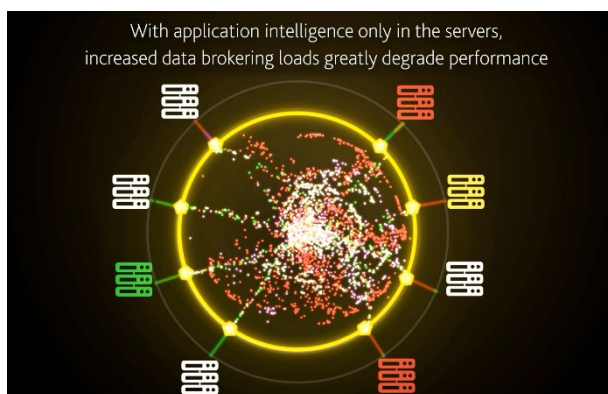
19. Xockets' New Cloud Fabric Patents, for example, further describe connecting together these DPUs in a novel way to form a new cloud network fabric for brokering collective communication independent of the limitations of existing cloud networks. This new cloud fabric is designed for even faster, lower-cost collective communication among server processors, and for in-network computing operations such as sorting, organizing, and reducing/combining data-intensive workloads in distributed computing. This new cloud fabric enables the training of large AI models across GPUs in a matter of weeks or months rather than many years as would otherwise be required. In this way, Xockets' DPU inventions ensure that training large models for AI and the production of AI can be made widely available and affordable to every business in every industry to drive forward a new industrial revolution.

## **II. XOCKETS PRESENTED ITS PATENTED DPU TECHNOLOGY, WHICH WAS THEN STOLEN BY DEFENDANTS**

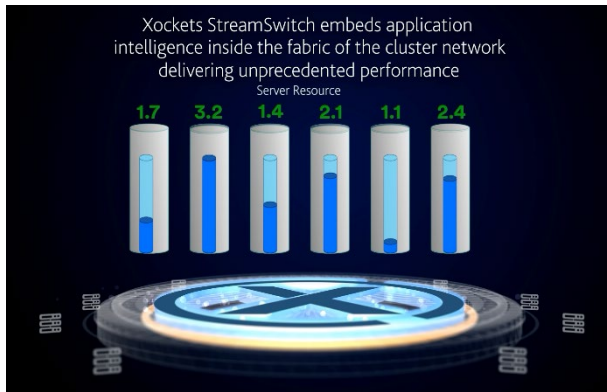
20. Xockets developed the world's first DPUs in a product it called the StreamSwitch. Xockets publicly displayed its patented DPU architecture in the StreamSwitch at Strata, the industry's premier big data and network technology conference, in the Fall of 2015.



21. At the Conference, Xockets demonstrated its revolutionary new DPU computing architecture and the “unprecedented performance” benefits it provides by offloading, accelerating, and isolating processing of data-intensive workloads from server processors and conventional computing, and by forming a new cloud network fabric of interconnected DPUs that can operate independent of the performance limitations of existing cloud networks:







## The result:

- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure

22. Xockets presented its DPU technology to Microsoft in 2016, demonstrating the invention's ability to accelerate computing performance on cloud data-intensive workloads—including “Big Data,” “Machine Learning” (which includes AI), “Security,” and “Encryption / Decryption” workloads—thousands of times faster using a fraction of the resources:

### WHAT DOES XOCKETS DO?

**XOCKETS DESIGNS THE XSTREAM APPLIANCE**

Public cloud providers, web-scale services companies, and OEMs can directly create new, unique, and powerful **hardware-accelerated services, just by programming software.**

*How?*

The XStream contains the worlds first physical, streaming processors. Our appliance inserts stream processing into the spine of clusters making the most difficult Machine Learning, batch Map-Reduce, or in-memory streaming analytics applications thousands of times faster, using a fraction of resources.

CONFIDENTIAL AND PROPRIETARY

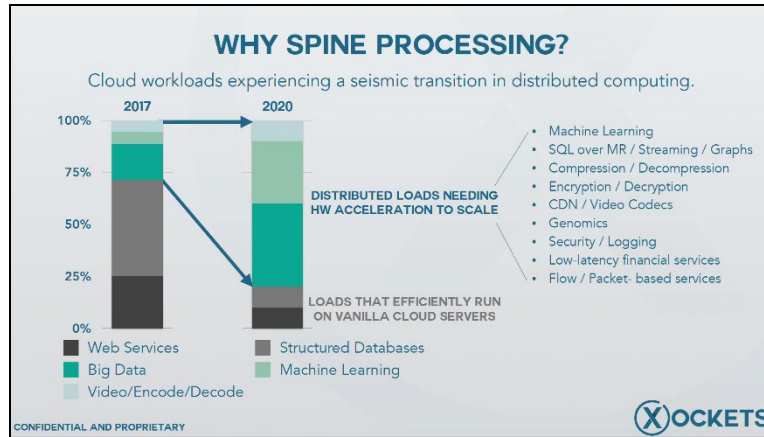
### XSTREAM APPLIANCE

320 Gb/s to 2.2 Tb/s of streaming processing

- >1000x Faster BigData computing
- >1000x Faster BigData repartitioning / sort
- >1000x Faster database joins
- >10x ROI in Machine learning over GPUs
- Less than 2x cost of server
- No change to users' code
- Available for Hadoop and Spark demonstrations today

**TOP OF RACK, BUMP-IN-WIRE DEPLOYMENT**  
XStream inserts reconfigurable, streaming processors into the switching spine of clusters

CONFIDENTIAL AND PROPRIETARY



23. Xockets' DPU technology was thereafter adopted by Mellanox in 2016<sup>7</sup> without Xockets' knowledge or permission for cloud offload use of server processors by Microsoft and other customers.

24. Mellanox's use of Xockets' DPU technology led to NVIDIA later acquiring Mellanox in order to drive forward NVIDIA's collaborations with Microsoft to dominate the markets for AI equipment and services using Xockets' DPU technology in its products. These collaborations continue to this day.

25. Thus, instead of licensing the technology, both Microsoft and NVIDIA chose to stand on the backs of Xockets' and Dr. Dalal's ingenuity, hard work, and innovations, and attempted to shut Xockets out of the market by misappropriating Xockets' technology and patents, all while NVIDIA was falsely proclaiming itself as the pioneer of accelerated computing and AI with its DPUs.

26. In addition, NVIDIA and Microsoft chose to form a cartel that leveraged Xockets' DPU technology to create the dominant market position these two companies hold today.

<sup>7</sup> <https://www.businesswire.com/news/home/20160615005424/en/Mellanox-Announces-ConnectX-5-the-Next-Generation-of-100G-InfiniBand-and-Ethernet-Smart-Interconnect-Adapter>; <https://www.servethehome.com/mellanox-connectx-5-vpi-100gbe-and-cdr-ib-review/mellanox-connectx-4-connectx-5-and-connectx-6-ethernet-comparison-chart-1>.



27. Microsoft has control over the leading generative artificial intelligence models in the world. Microsoft and NVIDIA have formed a cartel through an extensive series of what they euphemistically refer to as “collaborations” to maintain or create a monopoly in GPU-enabled generative artificial intelligence.

### III. XOCKETS’ DPU INVENTIONS ARE ESSENTIAL ELEMENTS OF NVIDIA’S AND MICROSOFT’S GPU-ENABLED SERVER COMPUTER SYSTEMS

28. NVIDIA’s systems feature distinct DPUs, including BlueField and ConnectX DPUs, as well as NVLink Switch DPUs, that offload key data-intensive workloads from server processors, including, among others, security, networking, and storage operations, and that form a virtual switching fabric for accelerating collective communication among server processors and enabling in-network computing of data-intensive workloads in training machine learning/artificial intelligence (ML/AI) models, all as claimed in the Xockets Patents.

29. For example, NVIDIA publicly states that “[t]he best definition of the DPU’s mission is to offload, accelerate, and isolate infrastructure workloads” and further explains each function<sup>8</sup>:

- **Offload:** Take over infrastructure tasks from the server CPU so more CPU power can be used to run applications.
- **Accelerate:** Run infrastructure functions more quickly than the CPU can, using hardware acceleration in the DPU silicon.
- **Isolate:** Move key data plane and control plane functions to a separate domain on the DPU, both to relieve the server CPU from the work and to protect the functions in case the CPU or its software are compromised.

A DPU should be able to do all three tasks.

---

<sup>8</sup> <https://developer.nvidia.com/blog/offloading-and-isolating-data-center-workloads-with-bluefield-dpu>.

30. Microsoft is a customer of NVIDIA and with privileged access to NVIDIA's infringing GPU-enabled server computer systems and components for AI, which Microsoft uses in, inter alia, its Microsoft Azure Cloud computing platform.

31. Microsoft is combining its privileged access to NVIDIA's equipment and its control over the leading generative artificial intelligence models in the world to leverage NVIDIA's monopoly over AI equipment to create a monopoly in AI platforms based on the equipment. Microsoft and NVIDIA have formed a cartel to affect this process.

32. This case focuses on NVIDIA's and Microsoft's infringement of Xockets' inventions, including in GPU-enabled server computer systems that use Xockets' claimed DPU computing architecture (e.g., enabled by NVIDIA's BlueField and ConnectX DPUs) as well as Xockets' claimed DPU cloud network fabric architecture (e.g., enabled by NVIDIA's NVLink Switch DPUs). The infringing systems include, for example, NVIDIA's existing Hopper GPU-enabled server computer systems for AI and the greatly expanded infringement in NVIDIA's upcoming Blackwell GPU-enabled server computer systems for AI scheduled for release this Fall 2024.

33. NVIDIA has been making and selling DPUs for its cloud GPU-enabled server systems since at least as of April 2020, when it completed its purchase of Mellanox (for its Bluefield and ConnectX DPUs) for approximately \$7 billion. NVIDIA's founder and CEO, Jensen Huang, declared NVIDIA's acquisition of Mellanox "a homerun deal":

"This is a homerun deal. Man I've been dreaming about this. You know the most important computer today is the data center, it is the epicenter of the computer industry. And *the most important applications that run in the data center today are AI applications and Big Data analytics applications. Doing computation on artificial intelligence... and moving huge amounts of data around is what drives the data center architectures today.* And so we are combining the leaders of AI computing and high speed

networking and data processing into one company. This is really quite extraordinary.”<sup>9</sup>

34. Explaining the Mellanox acquisition, Huang also stated:

“We believe that in future datacenters, the compute will not start and end at the server, but ***the compute will extend into the network. And the network itself, the fabric, will become part of the computing fabric.***”<sup>10</sup>

35. Huang explained the significance of DPU technology:

***[W]hen you take a large scale problem that spans the whole datacenter – it doesn’t fit in a single computer – and you accelerate the computation by several orders of magnitude . . . then the network becomes the problem, and it needs to be very fast. And so that’s the reason our relationship with Mellanox goes back a decade and we’ve been working with them for quite a long time. The networking problem is much, much more complex than just having faster and faster networking. And the reason for that is because of the amount of data that you are transmitting, synchronizing, collecting, and reducing across this distributed data center-scale computer and the computation on the fabric itself is complicated. . . . Putting intelligence in the network – and processing in the network – is vitally important to performance.***<sup>11</sup>

36. Huang also described the importance of offloading to a DPU, just as disclosed in the Xockets Patents:

A lot of datacenters today have every single packet that is transmitted secured because you want to reduce the attack surface of the datacenter until it’s basically every single transaction. There’s no way you going to do that on the CPU. ***So you have to move the networking stack off. You want to move the security stack off and you want to move the data processing and data movement stack off.*** And this is something that you want to do right at the NIC before it even comes into the computer and at the NIC before it leaves the computer. The onion, celery, and carrots – you know, ***the holy***

<sup>9</sup> <https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html>. All emphases are added unless otherwise indicated.

<sup>10</sup> <https://www.hpcwire.com/2019/03/14/why-nvidia-bought-mellanox-future-datacenters-will-belike-high-performance-computers>.

<sup>11</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

*trinity of computing* soup – is the CPU, the GPU, and the DPU. . . . *A DPU is going to be programmable, it's going to do all of that processing that you and I have already talked about, and it's going to offload the movement of data into the granular processing of the data as it's being transmitted and keep it from ever bothering the CPUs and GPUs* and avoid redundant copies of data. *That's the architecture of the future.* And that's the reason why we're so excited about Mellanox.<sup>12</sup>

37. Further explaining the importance of Xockets' revolutionary DPU architecture, Huang declared:

One of the most important things to disaggregate out of the server node and its CPU is the data processing. That is a giant amount of unnecessary CPU cores running unnecessary software in the datacenter. I don't know how much – its maybe 30 percent to 50 percent. . . . *I really do think that when you offload [to] the data processing on the SmartNIC*, when you're able to disaggregate the converged server, *when you can put accelerators anywhere in datacenter* and then can compose and reconfigure that datacenter for this specific workload – *that's a revolution.*<sup>13</sup>

38. NVIDIA explained that “DPUs are an essential element of modern and secure data centers in which CPUs, GPUs and DPUs are able to combine into a single computing unit that's fully programmable, AI-enabled and can deliver levels of security and compute power not previously possible.”<sup>14</sup> Similarly, in an April 2021 press release, NVIDIA stated that “[a] new type of processor, designed to process data center infrastructure software, is needed to offload and accelerate the tremendous compute load of virtualization, networking, storage, security and other

---

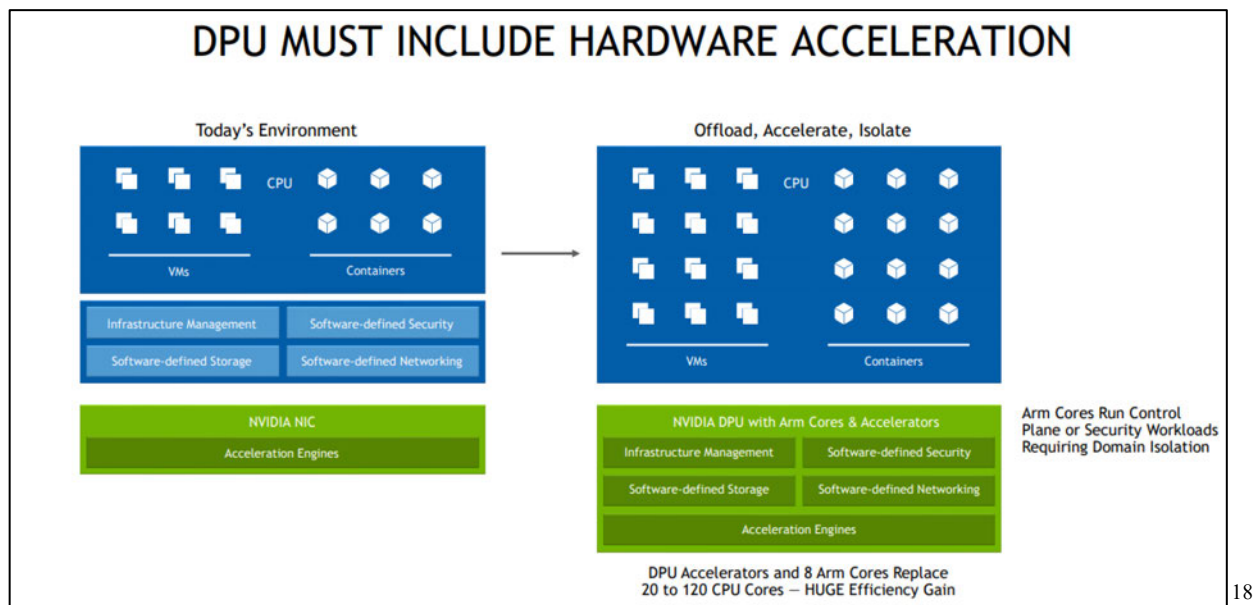
<sup>12</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

<sup>13</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

<sup>14</sup> <https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center>.

cloud-native AI services. The time for BlueField DPU has come.”<sup>15</sup> NVIDIA has described the DPU as a “new pillar” that is “designed to offload, accelerate, and isolate infrastructure workloads and bring efficiency and security to software defined workloads such as networking security and storage while freeing CPU resources by up to 30%.”<sup>16</sup>

39. In fact, NVIDIA illustrates how “[o]ffloading infrastructure tasks to the DPU improves server performance, efficiency, and security”<sup>17</sup> using Xockets’ DPU computing architecture:



18

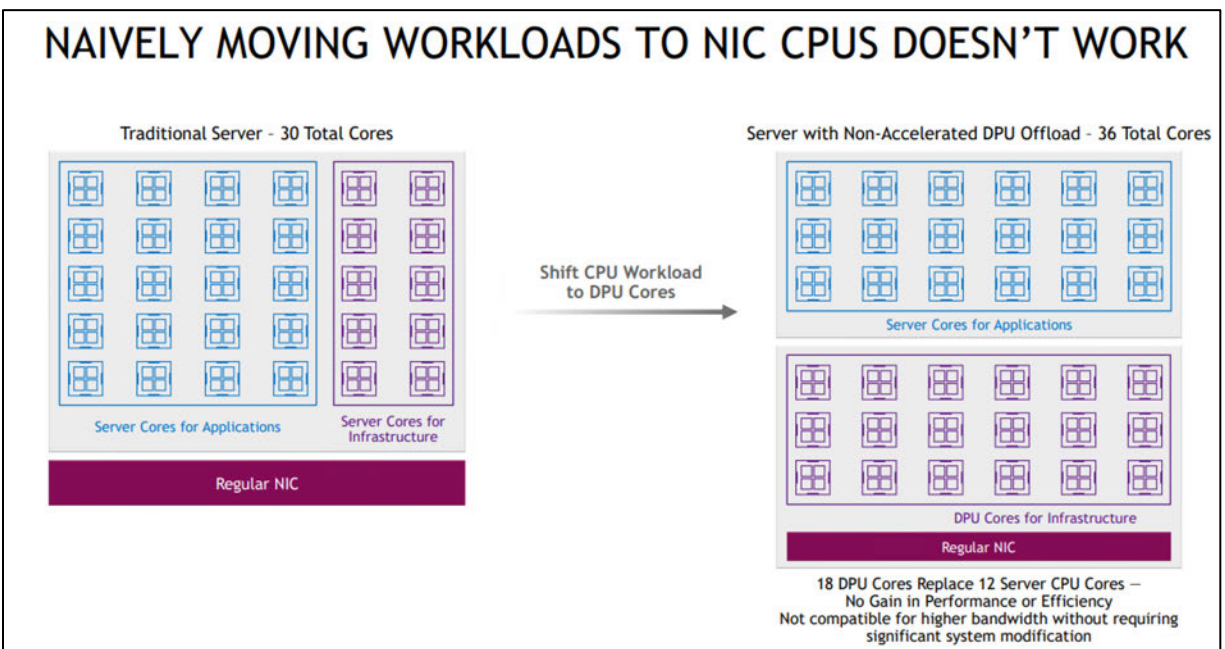
<sup>15</sup> <https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3>.

<sup>16</sup> NVIDIA DOCA Software Framework, <https://www.youtube.com/watch?v=htR19rdBicA>.

<sup>17</sup> <https://developer.nvidia.com/blog/offloading-and-isolating-data-center-workloads-with-bluefield-dpu>.

40. NVIDIA calls this a “fundamental new architecture.”<sup>19</sup> NVIDIA’s CEO Huang refers to the DPU paradigm as a “fundamental transition” necessitated by the fact that “CPU scaling [Moore’s law] has ended. We need a new computing approach and accelerated computing is the path forward. . . . This way of doing computation is a reinvention from the ground up.”<sup>20</sup>

41. Further, NVIDIA illustrates that Xockets’ DPU computing architecture is a breakthrough innovation, admitting that “naively moving workloads to NIC CPUs doesn’t work”<sup>21</sup>:



<sup>19</sup> <https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3> (“Modern hyperscale clouds are driving a fundamental new architecture for data centers,” said Jensen Huang, founder and CEO of NVIDIA.”).

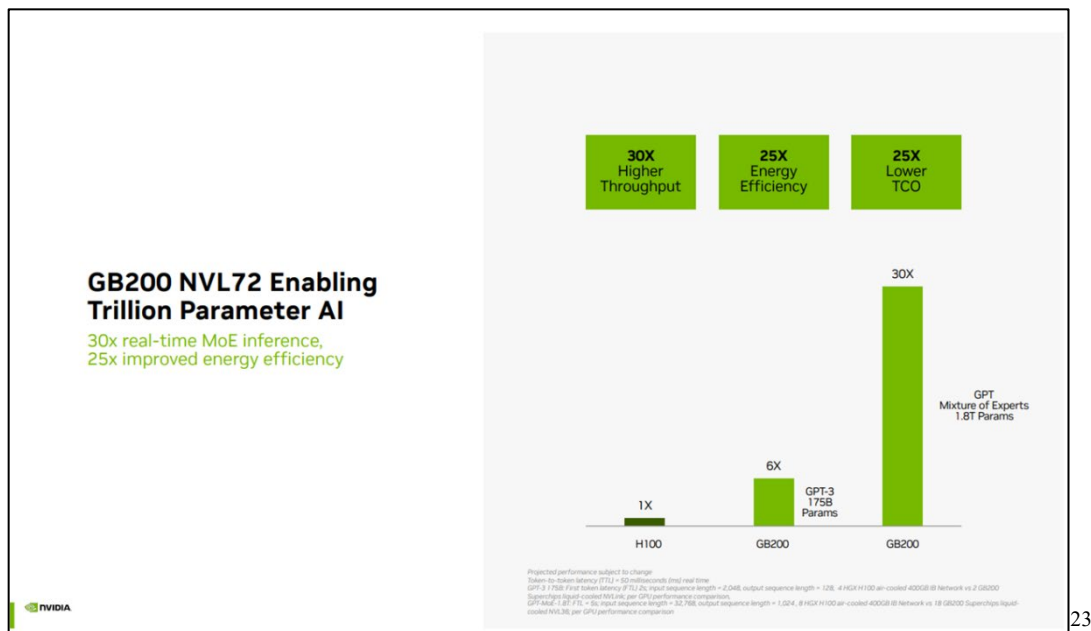
<sup>20</sup> [https://www.nvidia.com/en-us/events/computex/?nvid=nv-int-cwmfg-130532#cid=cmptx23e\\_nv-int-cwmfg\\_en-us](https://www.nvidia.com/en-us/events/computex/?nvid=nv-int-cwmfg-130532#cid=cmptx23e_nv-int-cwmfg_en-us); see also NVIDIA Keynote at COMPUTEX 2023, <https://www.youtube.com/watch?v=i-wpzs9ZsCs&t=875s> (14:35–15:47).

<sup>21</sup>

<https://hc33.hotchips.org/assets/program/conference/day1/HC2021.NVIDIA.IdanBurstein.v08.no.recording.pdf>.

42. While in the media NVIDIA and Huang claimed this technology as NVIDIA's own, Xockets invented, developed, patented, and presented to the industry this technology years earlier.

43. After first learning of NVIDIA's infringement and its collaborations with Microsoft, Dr. Dalal, personally provided NVIDIA with notice of the Xockets Patents on February 10, 2022. NVIDIA did not cease its infringing conduct, or even seek to negotiate for rights to the Xockets Patents. Instead, NVIDIA doubled-down on its infringing conduct by ceasing further contact with Dr. Dalal and exponentially expanding its infringement with the release of its NVLink Switch DPUs; to accelerate the training of large models for AI, including in extraordinary ways with its Blackwell GPU-enabled server computer systems to be released this Fall 2024.<sup>22</sup>



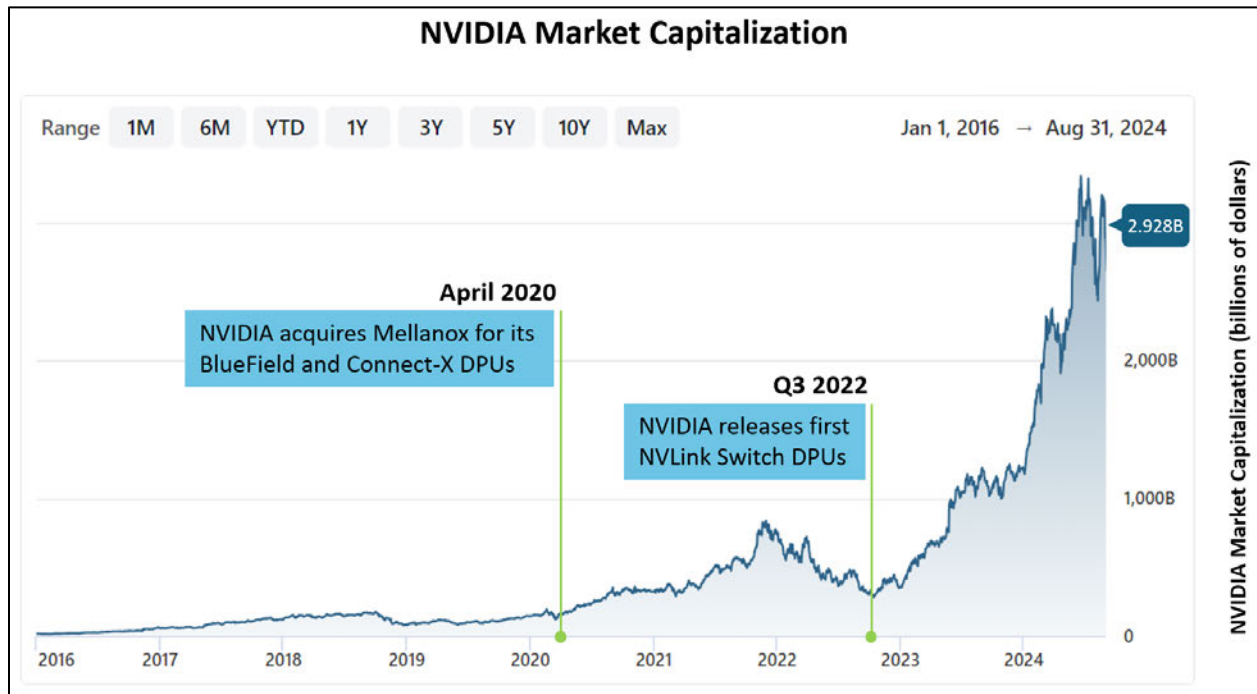
23

<sup>22</sup> <https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure> (“NVIDIA DGX H100 systems, DGX PODs and DGX SuperPODs will be available from NVIDIA's global partners starting in the third quarter [of 2022].”); <https://developer.nvidia.com/blog/?p=53977>.

<sup>23</sup> [https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024\\_page\\_26](https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_26).



44. Notably, since the launch of its systems that use Xockets' DPU technology, NVIDIA's market capitalization has surged from approximately \$180 billion in April 2020 (when NVIDIA acquired Mellanox and released its BlueField and ConnectX DPUs) to approximately \$3 *trillion* as of the end of August 2024 (following the release of its NVLink DPUs)<sup>24</sup>:



45. Huang and NVIDIA's own admissions, examples of which are above, demonstrate that NVIDIA vaulted to become one of the world's most valuable corporations by market capitalization in meaningful part through its widespread implementation and deployment of Xockets' patented DPU technology, described in and protected by Xockets' New Cloud Processor Patents and New Cloud Fabric Patents.

<sup>24</sup> <https://stockanalysis.com/stocks/nvda/market-cap> (Aug. 2024).



#### IV. NVIDIA’S AND MICROSOFT’S PREDATORY INFRINGEMENT THREATENS XOCKETS AND ITS EXCLUSIVE PATENT RIGHTS

46. Defendants NVIDIA and Microsoft have engaged and continue to engage in rampant infringement of Xockets’ patents, with the expectation that in nearly all cases “David” will not have the wherewithal to take on “Goliath” or withstand a war of attrition, as is consistent with the practice of “efficient infringement” (also known as “predatory infringement”) as described in more detail below. This business model, which is practiced by the largest technology companies (or Big Tech), looks to hand off the IP theft to lawyers to “clean the mess up” later.<sup>25</sup> In this case, Xockets’ patented architecture was misappropriated without permission and its exclusive rights violated, and Xockets was unable to raise sufficient capital to continue its hardware business.

47. As described earlier, the United States Constitution grants Congress the express power to promote the “useful Arts” by “securing for limited Times to [] Inventors the exclusive Right to their [] Discoveries.” As James Madison explained in The Federalist No. 43, “[t]he utility of this power [delegated to Congress to protect patents] will scarcely be questioned. . . . The public good fully coincides in both cases with the claims of individuals.” Recognizing the crucial role that the patent system would play in the fledgling country’s innovation economy, Congress wasted no time in exercising that power by enacting the first Patent Act in 1790 and granting these exclusive rights which continue to this day. Since then, the United States patent system has been a crown jewel of our Republic and a gold standard among patent systems, creating incentives that brought the United States to world leadership in technology and innovation. However, Big Tech

---

<sup>25</sup> <https://www.theverge.com/2024/8/14/24220658/google-eric-schmidt-stanford-talk-ai-startups-openai> (describing former Google CEO Eric Schmidt’s discussion of this practice during a recent recorded talk at Stanford University, which was taken down from the University’s YouTube channel due to public outrage).

companies like NVIDIA have lobbied all branches of our government for more than a decade to weaken these exclusive property rights in the hands of the people.

48. To this day, economists have endorsed the Founders’ recognition of the importance of incentivizing innovation through exclusive property rights. For example, leading economists Kevin A. Hassett, an economic advisor to Presidents Bush and Trump, and Robert J. Shapiro, an economic advisor to Presidents Clinton, Obama, and Biden, both agree on and emphasize the central role that innovation plays in a nation’s economy: ***“Innovation is widely recognized by economists as the most powerful factor that can drive changes in an economy’s underlying rates of productivity and growth.”*** The quality of the new ideas embodied in those innovations and the pace at which innovations are developed and applied, therefore, significantly affect a nation’s prosperity.” And they emphasize that ***“One legal aspect is especially critical to the development and broad application of economically-powerful ideas – the strict protection and enforcement of intellectual property rights,”*** as the Founders enshrined in the Constitution.<sup>26</sup>

49. Yet, today, the world’s largest companies in particular engage in widespread and willful infringement. They call this “efficient infringement,” but innovators who are the victims of these actions call it “predatory infringement.”

50. As the Hudson Institute, an influential non-partisan policy research institute, explains in a 2024 study, injunctive relief is critical to stopping predatory infringement:

***[W]here an injunction is not available, parties who wish to use the technology have no need to negotiate in good faith—or at all. These parties can coerce a “compulsory license” to use the technology through predatory infringement, having only to accept whatever “price” is set after the fact by courts or regulators. . . .***

---

<sup>26</sup> Kevin A. Hassett & Robert J. Shapiro, What Ideas Are Worth: The Value of Intellectual Capital and Intangible Assets in the American Economy, available at [https://www.sonecon.com/docs/studies/Value\\_of\\_Intellectual\\_Capital\\_in\\_American\\_Economy.pdf](https://www.sonecon.com/docs/studies/Value_of_Intellectual_Capital_in_American_Economy.pdf).

*When property owners lose the ability to say “no” when faced with the unauthorized access or use of their property, common sense says that infringement will become more common and commercial transactions more difficult and protracted. In patent law, this is now known as “predatory infringement,”* in which defendants choose a commercial strategy of “infringe now, pay later,” at worst, or, at best, they get away with infringement through a lengthy legal battle of attrition in which the patent owner ultimately just gives up.<sup>27</sup>

51. The erosion of exclusivity from the United States patent system caused by the practice of “predatory infringement” by Big Tech companies has had a measurable detrimental impact on the innovation economy in the United States. Once the world leader in technology innovation, the United States by one account now trails China in 37 of 44 critical technology areas.<sup>28</sup>

52. Accordingly, Xockets not only seeks damages for Defendants’ past infringement, it seeks to enjoin Defendants’ future infringement, including the sale of all NVIDIA Accused Products and all Microsoft’s use of the NVIDIA Accused Products as a remedy for their willful and unlawful behavior, including enjoining the release of NVIDIA’s new Blackwell GPU-enabled server computer systems and Microsoft’s use thereof for generative AI.

## **THE PARTIES**

### **I. XOCKETS**

53. Xockets is a Texas corporation, with its principal place of business located in the Temple Office Park at 2027 South 61st Street, Suite 107, Temple, Texas 76504. Temple’s thriving

---

<sup>27</sup> Kristen J. Osenga, The Loss of Injunctions Under *eBay*: Evidence of the Negative Impact on the Innovation Economy, Hudson Institute (Feb. 28, 2024), available at <https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impact-innovation-economy>.

<sup>28</sup> See <https://www.aspi.org.au/report/critical-technology-tracker>.

health and science industry<sup>29</sup> is an environment that Xockets believes will facilitate development of applications of DPU technology and ML/AI to improve people's health and lives.

54. Xockets is the developer and owner of foundational technology used to make today's AI breakthroughs possible.

55. Xockets was founded in 2012 by Dr. Parin Dalal.

56. Funding for Xockets was provided by notable investors, including the current CTO of Intel, Dr. Greg Lavender, who was a long-time Xockets Board member, Robert Cote, one of the nation's leading IP investors and lawyers, and also a Xockets Board member, and Jerry Yang, cofounder of Yahoo.

57. Xockets specializes in the development of infrastructure products for distributed computing within the technology sector, including the integration of hardware and software acceleration into appliances for distributed computing, including AI. Its developments are designed to enhance the performance of open source frameworks, reduce power consumption, and lower capital costs, all while being compatible with commodity scale-out data centers.

58. Xockets is the assignee and owns all right, title, and interest to the New Cloud Processor Patents (i.e., the '209 Patent – DPU Computing Architecture, Security; '924 Patent – DPU Network Overlay, Security; and '350 Patent – DPU Stream Processing) and the New Cloud Fabric Patents (i.e., the '297 Patent – DPU Cloud Network Fabric; '161 Patent – DPU Cloud Network Fabric; '092 Patent – DPU In-Network Computing; and '640 Patent – DPU In-Network Computing).

---

<sup>29</sup> See <https://templeedc.com/why-temple-tx-offers-a-healthy-ecosystem-for-healthcare-businesses>.

## **II. NVIDIA**

59. Defendant NVIDIA is a Delaware corporation. NVIDIA is registered with the State of Texas and may be served with process through its registered agent, Corporation Service Company d/b/a CSC-Lawyers Incorporating Service Company, 211 E. 7th Street, Suite 620, Austin, Texas 78701. NVIDIA maintains a facility in Austin at 11001 Lakeline Boulevard, Suite #100 Building 2, Austin, Texas 78717.

## **III. MICROSOFT**

60. Defendant Microsoft Corporation is a Washington corporation. Microsoft is registered with the State of Texas and may be served with process through its registered agent, Corporation Service Company d/b/a CSC-Lawyers Incorporating Service Company, 211 E. 7th Street, Suite 620, Austin, Texas 78701. Microsoft maintains a facility in Austin at 10900 Stonelake Blvd, Suite 225, Austin, Texas 78759.

## **IV. RPX**

61. Defendant RPX Corporation is a Delaware corporation. RPX is registered with the State of Texas and may be served with process through its registered agent, Incorporating Services, Ltd., 3610-2 North Josey, Suite 223, Carrollton, Texas 75007.

## **JURISDICTION AND VENUE**

62. This Court has subject matter jurisdiction over federal antitrust claims pursuant to 15 U.S.C. §§ 15 and 26 and 28 U.S.C. § 1331.

63. This Court has subject matter jurisdiction over patent infringement claims pursuant to 28 U.S.C. §§ 1331 and 1338, as those claims arise under the patent laws of the United States (35 U.S.C. §§ 1 *et seq.*).

**I. NVIDIA**

64. NVIDIA is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over NVIDIA because NVIDIA has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western District of Texas. For example, on information and belief, NVIDIA has committed, and continues to commit, violations of federal antitrust laws in the State of Texas and this District; NVIDIA has committed, and continues to commit, the tort of patent infringement in the State of Texas and this District; NVIDIA purposefully availed itself of the privileges of conducting business in the State of Texas and this District; and NVIDIA regularly conducts and solicits business within the State of Texas and this District.

65. Personal jurisdiction also exists over NVIDIA because NVIDIA, directly or through subsidiaries, makes, uses, sells, offers for sale, imports, advertises, makes available, and/or markets products and/or services within Texas, including in the Western District of Texas, that infringe one or more claims of the Asserted Patents. Further, on information and belief, NVIDIA has placed or contributed to placing infringing products and/or services into the stream of commerce knowing or understanding that such products and/or services would be sold and used in this District.

66. Furthermore, personal jurisdiction over NVIDIA in this action comports with due process. For example, NVIDIA has conducted and regularly conducts business within this District; NVIDIA has purposefully availed itself of the privileges of conducting business in this District; and NVIDIA has sought protection and benefit from the laws of the State of Texas by placing infringing products into the stream of commerce through an established distribution channel with the awareness and/or intent that they will be purchased by consumers in this District. Having

purposefully availed itself of the privilege of conducting business within this District, NVIDIA should reasonably and fairly anticipate being brought into court here.

67. NVIDIA has repeatedly acknowledged this Court has personal jurisdiction over it. *See, e.g., Vantage Micro LLC v. NVIDIA Corporation*, Case No. 6:19-cv-00582-RP, Dkt. 22 (W.D. Tex., Jan. 4, 2020) (admitting to personal jurisdiction); *Ocean Semiconductor LLC v. NVIDIA Corporation*, Case No. 6:20-cv-01211-ADA, Dkt. 14 (W.D. Tex., Mar. 12, 2021) (same). Further, NVIDIA has admitted “it is subject to this Court’s general personal jurisdiction.” *Id.*

68. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because NVIDIA has a regular and established place of business in this District through which it transacts business and where it may be found. Further, as detailed herein, NVIDIA is subject to this Court’s personal jurisdiction with respect to the violations described herein.

69. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(c) and/or 1400(b), including but not limited to because NVIDIA has committed acts of infringement in this District and has a regular and established place of business in this District. For example, NVIDIA maintains an office in this District, including at 11001 Lakeline Blvd., Suite #100 Bldg. 2, Austin, Texas 78717. Further, by way of example and without limitation, NVIDIA makes, uses, sells, offers to sell, and/or imports products and/or services that are accused of infringing the Asserted Patents into and/or within this District and maintains a permanent and/or continuing presence within this District.

70. On information and belief, NVIDIA designs, develops, manufactures, sells, and offers to sell the Accused Products in this District. For example, as described in detail below, NVIDIA acquired Mellanox in April 2020. On information and belief, as part of this acquisition,

NVIDIA inherited Mellanox's design and development teams for the Accused Products located in Austin. As another example, on information and belief, NVIDIA designs and develops its NVLink Cloud Network Fabric, including but not limited to its NVLink Switch DPUs and NVLink Cabling, in this District. As another example, on information and belief, NVIDIA's sales team for the Accused Products is based in this District.

71. Further, on information and belief, NVIDIA employs engineers in this District to design and develop products, devices, systems, and/or components of systems that are accused of infringing one or more claims of the Asserted Patents. For example, as of the date of this Complaint, NVIDIA has several open job postings for its Austin office for engineers relating to the Accused Products.<sup>30</sup>

72. In addition, on information and belief, NVIDIA has not disputed that venue is proper in this District in cases filed against it in this District. *See, e.g., Vantage Micro LLC v. NVIDIA Corp.*, No. 6:19-cv-00582, Dkt. No. 22; *Polaris Innovations Ltd. v. Dell Inc. et al.*, No. 5:16-cv-00451, Dkt. No. 19; *Cirrus Logic, Inc. v. ATI Techs., et al.*, No. 1:03-cv-00302, Dkt. No. 6.

## II. MICROSOFT

73. Microsoft is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over Microsoft because Microsoft has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western District of Texas. For example, on information and belief, Microsoft has committed, and continues to commit, violations of federal antitrust laws in the State of Texas and

---

<sup>30</sup> *See* <https://nvidia.wd5.myworkdayjobs.com/NVIDIAExternalCareerSite?locations=91336993fab910af6d702b631b94c2de>.



this District; Microsoft has committed, and continues to commit, the tort of patent infringement in the State of Texas and this District; Microsoft purposefully availed itself of the privileges of conducting business in the State of Texas and this District; and Microsoft regularly conducts and solicits business within the State of Texas and this District.

74. Personal jurisdiction also exists over Microsoft because Microsoft, directly or through subsidiaries, makes, uses, sells, offers for sale, imports, advertises, makes available, and/or markets infringing products and/or services within Texas, including in the Western District of Texas, that infringe one or more claims of the Asserted Patents. Further, on information and belief, Microsoft has placed or contributed to placing infringing products and/or services into the stream of commerce knowing or understanding that such products and/or services would be sold and used in this District.

75. In addition, on information and belief, Microsoft has not disputed personal jurisdiction in cases filed against it in this District. *See, e.g., Panther Innovations v. Microsoft Corp.*, No. 6:20-cv-01071, Dkt. No. 14; *Exafer Ltd v. Microsoft Corp.*, No. 1:20-cv-00131, Dkt. No 15; *WSOU Investments, LLC v. Microsoft Corp.*, No. 6:20-cv-00464, Dkt. No. 20; *Zeroclick, LLC v. Microsoft Corp.*, No. 1:20-cv-00272, Dkt. No. 14.

76. Furthermore, personal jurisdiction over Microsoft in this action comports with due process. For example, on information and belief, Microsoft has conducted and regularly conducts business within this District; Microsoft has purposefully availed itself of the privileges of conducting business in this District; and Microsoft has sought protection and benefit from the laws of the State of Texas by placing infringing products into the stream of commerce through an established distribution channel with the awareness and/or intent that they will be purchased by consumers in this District. Having purposefully availed itself of the privilege of conducting

business within this District, Microsoft should reasonably and fairly anticipate being brought into court here.

77. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because Microsoft has a regular and established place of business in this District through which it transacts business and where it may be found. Further, as detailed herein, Microsoft is subject to this Court's personal jurisdiction with respect to the violations described herein.

78. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(c) and/or 1400(b), including but not limited to because Microsoft has committed acts of infringement in this District and has a regular and established place of business in this District. For example, Microsoft maintains corporate offices in this District, including at 10900 Stonelake Boulevard, Suite 225, Austin, Texas, 78759 and Concord Park II, 401 East Sonterra Boulevard, Suite 300, San Antonio, Texas 78258.

79. In addition, on information and belief, Microsoft has not disputed that venue is proper in this District in cases filed against it in this District. *See, e.g., Panther Innovations v. Microsoft Corp.*, No. 6:20-cv-01071, Dkt. No. 14; *Exafer Ltd v. Microsoft Corp.*, No. 1:20-cv-00131, Dkt. No 15; *WSOU Investments, LLC v. Microsoft Corp.*, No. 6:20-cv-00464, Dkt. No. 20; *Zeroclick, LLC v. Microsoft Corp.*, No. 1:20-cv-00272, Dkt. No. 14.

### **III. RPX**

80. RPX is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over RPX because RPX has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western District of Texas. For example, on information and belief, RPX has committed, and continues to

commit, violations of federal antitrust laws in the State of Texas and this District; RPX purposefully availed itself of the privileges of conducting business in the State of Texas and this District; and RPX regularly conducts and solicits business within the State of Texas and this District.

81. Furthermore, personal jurisdiction over RPX in this action comports with due process. For example, on information and belief, RPX has conducted and regularly conducts business within this District; RPX has purposefully availed itself of the privileges of conducting business in this District; and RPX has sought protection and benefit from the laws of the State of Texas. Having purposefully availed itself of the privilege of conducting business within this District, RPX should reasonably and fairly anticipate being brought into court here.

82. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because RPX transacts business in this District. Further, as detailed herein, RPX is subject to this Court's personal jurisdiction with respect to the violations described herein.

### **JOINDER**

83. Joinder is proper under 35 U.S.C. § 299 because right to relief is asserted against the parties jointly, severally, or in the alternative with respect to or arising out of the same transaction, occurrence, or series of transactions or occurrences relating to the making, using, importing into the United States, offering for sale, or selling of the same accused product or process; and questions of fact common to all defendants or counterclaim defendants will arise in the action. For example, Microsoft infringes the asserted Xockets Patents at least through its use of the accused NVIDIA systems and components that NVIDIA sells to Microsoft, as alleged below.

## **FACTUAL ALLEGATIONS**

### **I. XOCKETS PATENTS**

84. Xockets' New Cloud Processor Patents (including the '209, '924, and '350 Patents) and its New Cloud Fabric Patents (including the '297, '161, '092, and '640 Patents) are exemplary Xockets patents that are targets of the illegal buyers' cartel discussed herein.

85. In addition, NVIDIA and Microsoft are infringing Xockets' New Cloud Processor Patents (including the '209, '924, and '350 Patents) and its New Cloud Fabric Patents (including the '297, '161, '092, and '640 Patents) by making, selling, and using GPU-enabled server computers for AI and delivering to customers AI servers with this equipment.

#### **A. THE NEW CLOUD PROCESSOR PATENTS**

86. Xockets invented the DPU and its virtual switch computing architecture years before the industry to enable accelerated computing and AI in cloud data centers, providing the versatility needed in offloading, accelerating, and isolating from cloud server processors (e.g., CPUs, GPUs, and hybrids of these server processors) the data-intensive computing tasks required to make distributed computing in data centers possible, including for training large models for AI.

87. Xockets' DPU computing architecture is protected by the New Cloud Processor Patents, including the '209, '924, and '350 Patents, discussed below.

##### **(1) The '209 Patent – DPU Computing Architecture, Security**

88. U.S. Patent No. 11,080,209 ("the '209 Patent") is entitled "Server Systems and Methods for Decrypting Data Packets With Computation Modules Insertable Into Servers That Operate Independent of Server Processors." The '209 Patent duly and legally issued on August 3, 2021, from U.S. Patent Application No. 15/396,334, filed on December 30, 2016.

89. The '209 Patent is a continuation of U.S. Patent Application No. 13/900,346, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The '209 Patent is entitled to the benefit of these earlier filed applications.

90. Xockets is the current owner of all rights, title, and interest in and to the '209 Patent, including the right to sue for past damages.

91. A true and correct copy of the '209 Patent is attached hereto as **Exhibit 1** and is incorporated by reference herein.

92. The '209 Patent relates to server systems in cloud data centers utilizing a novel computing architecture in a new cloud processor, or DPU, for offloading, accelerating, and isolating data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors), including for cloud infrastructure services and big data analytics applications such as those used in training large language models for AI. The server system can include a plurality of servers interconnected by a network. Each server includes a server processor that is configured to execute an operating system for the server. Each server further includes a computation module, or DPU, that is separate from the server processor and is coupled to the server processor by a bus. The computation module, or DPU, includes a virtual switch computing architecture for identifying and classifying packet flows, also known as sessions, and connecting together identified programmable logic pipelines of hardware accelerators or computation elements, comprising offload processors or offload processing circuits. The computation module, or DPU, uses packet data to define in the virtual switch the programmable logic pipelines to be formed for computational operations, or what is referred to as data-centric computing. This data-centric computing approach supports cloud computing at the speed of the network, or line rate.

The programmable hardware acceleration formed using the virtual switch is for performing computing operations on packet data independent of server processors in the cloud data center.

93. The invention of the '209 Patent solves a technological problem with prior art server systems and methods for performing computationally intensive workloads such as processing of packets for high-volume applications. For example, the '209 Patent explains that “[p]acket handling and security applications can require a significant amount of scarce computational resources in enterprise server or cloud based data systems.” '209 Patent, 1:27–29.

94. Conventional approaches of throwing more hardware at the problem under control of server processors, for processing data-intensive workloads and sources in cloud data centers, were too expensive and did not address the root cause of the problem, which was that the server processors in cloud servers were constantly interrupted and bottlenecked with data-intensive infrastructure services such as for “packet handling and transport services.” *Id.*, 1:32–38. These conventional architectures were thus ill-equipped to handle such high-volume applications: “Even idling, x86 processors use a significant amount of power, and near continuous operation for high bandwidth packet analysis functionality make the processor energy costs one of the dominant price factors.” *Id.*, 1:39–44. “In addition, issues with the high cost of context switching, the limited parallelism, and the security implications associated with running encryption/decryption modules on x86 processors have reduced the effectiveness of enterprise or cloud data security.” *Id.*, 1:44–48.

95. The '209 Patent improved upon the systems and methods in the prior art by introducing a new cloud processor located at the boundary of the network leading to each server processor, providing bump-in-the-wire hardware acceleration for data-intensive computing operations independent of server processors using the virtual switch computing architecture. As

described in the '209 Patent, this Xockets virtual switch is programmed to form programmable logic pipelines of hardware acceleration and has the versatility needed for cloud adoption. Such cloud processors or DPUs are referred to as computation modules or offload processing modules in the '209 Patent and in one embodiment are referred to as Xocket In-line Memory Modules ("XIMMs"). *See id.*, 2:13–20. "Using one or more XIMMs it is possible to execute lightweight packet handling tasks without intervention from a main server processor." *Id.*, 2:20–22.

96. The invention of the '209 Patent "can have high efficiency context switching, high parallelism, and can solve security problems associated with running encryption/decryption modules on x86 processors. Such systems as a whole are able to handle high network bandwidth traffic at a lower latency and at a very low power when compared to traditional high power 'brawny' server cores." *Id.*, 2:22–29. The invention of the '209 Patent can thus provide software-defined hardware acceleration in processing data-intensive workloads of cloud infrastructure services with lower power costs and high reliability. *See id.*, 2:29–33. A variety of cloud infrastructure services can thus be offloaded to the computation module and accelerated, and run independent of server processors at the line rate of the network, providing the versatility needed to offload and accelerate various cloud applications and infrastructure services "including but not limited to virtual private network (VPN) tunneling and signature detection and packet filtering as an intrusion prevention system (IPS)" such as used in cloud VPN communications, providing levels of compute and security in cloud data centers that were not previously possible. *See id.*, 2:46–50.

97. For example, Claim 18 of the '209 Patent is directed to:

18. A server system, comprising:

a plurality of servers interconnected by a network, each server including

a server processor configured to execute an operating system for the server,

at least one computation module, separate from the server processor and coupled to the server processor by at least one bus, the at least one computation module including

first processing circuits mounted on the computation module and configured to

execute header detection on packets received by the server,

classifying received packets by a session identifier, and

operate as a virtual switch to provide packets to circuits on the at least one computation module, and

at least decryption circuits implemented on programmable logic devices and configured to decrypt received packets; wherein

the computation modules execute header detection, classifying of packets, virtual switching of packets, and decryption of packets independent of the server processor of their respective server.

98. For example, Claim 20 of the '209 Patent is directed to:

20. The server system of claim 18, wherein the at least decryption circuits decrypt the received packets according to a virtual private network (vpn) encryption/decryption protocol.

99. NVIDIA is not licensed to the '209 Patent.

100. Microsoft is not licensed to the '209 Patent.

**(2) The '924 Patent – DPU Network Overlay, Security**

101. U.S. Patent No. 10,649,924 (“the '924 Patent”) is entitled “Network Overlay Systems and Methods Using Offload Processors.” The '924 Patent duly and legally issued on May 12, 2020, from U.S. Patent Application No. 15/396,323, filed on December 30, 2016.

102. The '924 Patent is a continuation of U.S. Patent Application No. 13/921,059, filed on June 18, 2013, and claims priority from U.S. Provisional Application Nos. 61/753,901;



61/753,906; 61/753,892, 61/753,899; 61/753,903; 61/753,895; 61/753,910; 61/753,904; and 61/753,907, all filed on January 17, 2013. The '924 Patent is entitled to the benefit of these earlier filed applications.

103. Xockets is the current owner of all rights, title, and interest in and to the '924 Patent, including the right to sue for past damages.

104. A true and correct copy of the '924 Patent is attached hereto as **Exhibit 2** and is incorporated by reference herein.

105. The '924 Patent relates to systems, hardware, and methods in cloud data centers utilizing a virtual switch computing architecture in a new cloud processor, or DPU, to offload, accelerate, and isolate data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors) to provide network overlay infrastructure services for improved cloud security, including by implementing network overlays for cloud VPN communications. In particular, the '924 Patent relates to network overlay services that are provided by the new cloud processor, or DPU, also called offload processor modules, that receives data packets and routes them to programmable logic pipelines of hardware accelerators comprising offload processors or offload processing circuits for packet encapsulation, decapsulation, modification, or data handling, such as in cloud VPN communications. The offload processor modules are mounted to a system bus of a host server, that further includes a host processor connected to the system bus. Offload processor modules include offload processors or offload processing circuits that function as hardware accelerators and are configured to encapsulate network packet data for transport on a logical network or decapsulate the network packet data received from the logical network. The offload processing circuits encapsulate or decapsulate network packet data independent of any host processor and provide programmable hardware

acceleration to packet encapsulation and decapsulation functions. This is critical to enabling cloud data centers to offload the provisioning for each customer a secure virtual network that is seemingly a different network to the customer, but is actually running on the same physical network.

106. The '924 Patent discloses and claims improved systems and methods for processing packets in cloud data centers using network overlay services. “Modern computing systems can support a variety of intercommunication protocols. In certain instances, computers can connect with each other using one network protocol, while appearing to outside users to use another network protocol. Commonly termed an ‘overlay’ network, such computer networks are effectively built on top of another computer network, with nodes in the overlay network being connected by virtual or logical links to the underlying network.” '924 Patent, 1:27–34.

107. While “[o]verlay networks are particularly useful for environments where different physical network servers, processors, and storage units are used, and network addresses to such devices may commonly change,” “overlay networks do require additional computational processing power to run.” *Id.*, 1:47–55. Therefore, “efficient network translation mechanisms are necessary, particularly when large numbers of network transactions occur.” *Id.*, 1:55–57.

108. To solve this issue, the '924 Patent provides improved systems, hardware, and methods that allow for “high speed and/or energy efficient processing of packet data that does not necessarily require access to computing resources of a host processor of a server, server rack system, or blade server.” *Id.*, 1:61–65. Instead, packets can be directed to and processed by offload processor modules, or DPUs, which can operate on the packet data independent of any host processors. Thus, using the invention of the '924 Patent, “server loads can be broken up across the offload processing cores and the host processing cores.” *Id.*, 5:19–21.

109. As a result, the invention of the '924 Patent “can provide improved computational performance as compared to traditional computing systems,” in providing Cloud VPN services, which “are often ill-equipped to handle such high volume applications.” *Id.*, 8:37–41.

110. For example, Claim 9 of the '924 Patent is directed to:

9. A method for providing network overlay services, comprising the steps of:

receiving network packet data from a data source in an offload processor module that is mounted to a system bus of a host server, the host server further including

at least one host processor connected to the system bus, and

a network interface device;

encapsulating the network packet data to create encapsulated network packets for transport on a logical network or decapsulating the network packet data to create decapsulated network packets for delivery to a network location, the encapsulating and decapsulating being executed by processing circuits mounted on the offload processor module and being executed independent of any host processor; and

transporting the encapsulated network packets or the decapsulated network packets out of the offload processor module; wherein

the logical network is overlaid on a physical network.

111. NVIDIA is not licensed to the '924 Patent.

112. Microsoft is not licensed to the '924 Patent.

### **(3) The '350 Patent – DPU Stream Processing**

113. U.S. Patent No. 11,082,350 (“the '350 Patent”) is entitled “Network Server Systems, Architectures, Components and Related Methods” for expanding the versatility of a DPU for cloud offload by adding general-purpose processors (e.g., ARM cores) with a modified architecture described in the '350 Patent that ensures that they can function as general-purpose hardware accelerators and process data-intensive workloads at the speed of the network, or line

rate. The '350 Patent duly and legally issued on August 3, 2021, from U.S. Patent Application No. 16/129,762, filed on September 12, 2018.

114. The '350 Patent is a continuation-in-part of U.S. Patent Application No. 15/396,318, filed on December 30, 2016, which is a continuation of U.S. Patent Application No. 13/900,318, filed on May 22, 2013, and U.S. Patent Application No. 15/283,287 ("the '287 Application"), filed on September 30, 2016. Further, the '287 Application is a continuation of International Patent Application Nos. PCT/US2015/023730 and PCT/US2015/023746, filed on March 31, 2015. In addition, the '350 Patent claims priority from U.S. Provisional Application Nos. 62/557,659; 62/557,661; 62/557,666; 62/557,670; 62/557,671; 62/557,675; 62/557,679; 62/557,687, all filed on September 12, 2017; U.S. Provisional Application No. 61/976,471, filed on April 7, 2014; U.S. Provisional Application Nos. 61/973,207 and 61/973,205, filed on March 31, 2014; U.S. Provisional Application Nos. 61/753,892; 61/753,895; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,910; 61/753,907; and 61/753,899, all filed on January 17, 2013; and U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The '350 Patent is entitled to the benefit of these earlier filed applications.

115. Xockets is the current owner of all rights, title, and interest in and to the '350 Patent, including the right to sue for past damages.

116. A true and correct copy of the '350 Patent is attached hereto as **Exhibit 3** and is incorporated by reference herein.

117. The '350 Patent relates to systems in cloud data centers utilizing a novel computing architecture in a new cloud processor, or DPU, also called hardware acceleration modules, to offload, accelerate, and isolate data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors). The system includes a server with a host processor

and at least one hardware acceleration (hwa) module, or DPU, physically separate from the host processor. Hardware accelerators or computing elements are formed in programmable logic pipelines on the hardware acceleration module that include offload processors or offload processing circuits configured to execute a plurality of processes, first memory circuits, and second memory circuits. The computing elements further include a hardware scheduler circuit for streaming network packet flows to the processing circuits (e.g., general-purpose processors such as ARM cores) using their associated memory circuits so that they function like hardware accelerators. The hardware acceleration module includes a data transfer fabric configured to enable data transfers between the processing circuits and the first and second memory circuits; wherein the computing elements are configured to transfer data to, or receive data from, any of: the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric. The addition of the hardware scheduler enables stream processing in the hardware acceleration modules, providing run-to-completion computational processing of packet flows in general-purpose processors that now function like hardware accelerators at the speed of the network, or line rate.

118. The invention of the '350 Patent solves a technological problem with prior art data processing systems, including systems and methods for processing large data sets. For example, the '350 Patent explains that “[c]onventional data intensive computing platforms for handling large volumes of unstructured data can use a parallel computing approach combining multiple processors and disks in large commodity computing clusters connected with high-speed communications switches and networks.” '350 Patent, 15:3–7. An exemplary programming model for processing large data sets is known as “map, reduce.” *Id.*, 15:14–17. However, in such

conventional systems, “data spills to disk are almost unavailable. This slows performance and such spilled data needs to be read back into server memory to continue processing.” *Id.*, 15:34–37.

119. Accordingly, the ’350 Patent recognized that “[i]t would be desirable to arrive at some way of increasing the performance of [] systems for processing unstructured data that do not suffer from the drawbacks of conventional approaches.” *Id.*, 15:43–46. To that end, the ’350 Patent discloses and claims improved systems and methods, including those “that can perform data processing, including ‘big’ data processing, by accelerating processing tasks with networked hardware accelerator (hwa) modules included in server systems.” *Id.*, 15:47–50. The improved systems “can provide map, reduce type processing, without data skew and/or spills to disk that can occur in conventional architectures.” *Id.*, 18:8–11. As an example, in the case of large language model training for AI requiring “machine learning applications to run across multiple computing elements on multiple networked servers,” the stream processing invention enables in-network or in-flight computing operations such as reduction/combining of training results. *Id.*, 9:12–14.

120. For example, Claim 1 of the ’350 Patent is directed to:

1. A device, comprising:

a server that includes a host processor and at least one hardware acceleration (hwa) module physically separate from the host processor and having

a network interface configured to virtualize functions by redirecting network packets to different addresses within the hwa,

at least one computing element formed thereon, the at least one computing element including

processing circuits configured to execute a plurality of processes including at least one virtualized function,

a scheduler circuit configured to allocate a priority to a processing of packets of one flow over those of another flow by the processing circuits,

first memory circuits,

second memory circuits, and

a data transfer fabric configured to enable data transfers between the processing circuits and the first and second memory circuits; wherein

the at least one computing element is configured to transfer data to, or receive data from, any of: the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric.

121. NVIDIA is not licensed to the '350 Patent.

122. Microsoft is not licensed to the '350 Patent.

**B. THE NEW CLOUD FABRIC PATENTS**

123. Xockets invented a DPU switching architecture for connecting together its DPUs in a novel way to form a “New Cloud Fabric” in data centers—one that can bypass the existing cloud network and its limitations to enable accelerated computing and AI in data centers, and turn every data center into an AI factory.

124. Xockets’ DPU switching architecture for forming a new cloud fabric is protected by the New Cloud Fabric Patents, including the '297 and '161 Patents, which claim this New Cloud Fabric, and the '092 and '640 Patents, which claim offloading from server processors, accelerating, and isolating the processing of data-intensive workloads in this DPU fabric.

125. Xockets’ New Cloud Fabric enables brokering of high-speed collective communication of data between server processors in a cloud data center and in-network computing to sort, organize, and reduce the data-intensive workloads involved in training AI models across GPU servers. This enables the higher speeds and power efficiency needed to make AI production affordable and widely available.

**(1) The '297 Patent – DPU Cloud Network Fabric**

126. U.S. Patent No. 10,223,297 (“the '297 Patent”) is entitled “Offloading of Computation for Servers Using Switching Plane Formed by Modules Inserted Within Such Servers” for forming a new cloud fabric that can operate independent of server processors. The '297 Patent duly and legally issued on March 5, 2019, from U.S. Patent Application No. 15/396,328, filed on December 30, 2016.

127. The '297 Patent is a continuation of U.S. Patent Application No. 13/900,222, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The '297 Patent is entitled to the benefit of these earlier filed applications.

128. Xockets is the current owner of all rights, title, and interest in and to the '297 Patent, including the right to sue for past damages.

129. A true and correct copy of the '297 Patent is attached hereto as **Exhibit 4** and is incorporated by reference herein.

130. The '297 Patent relates to server systems in cloud data centers, and more particularly to computation modules, or DPUs, also called offload processor modules, in such systems that are connected to form a new switching plane or new cloud fabric that can operate independent of server processors. The system includes a plurality of first server modules interconnected to one another via a communication network, wherein each first server module includes a first switch for forming a first switching plane, at least one main processor, and at least one computation module, or DPU, coupled to the main processor by a bus. Each computation module, or DPU, includes a second switch and a plurality of computation elements that function as hardware accelerators, comprising offload processors or offload processing circuits for performing programmable hardware acceleration in the network. The second switches of the first server modules form a second switching plane or cloud fabric for the ingress and egress of network



packets independent of any main processors of the first server modules. Furthermore, the second switches include virtual switches that switch together programmable logic pipelines of hardware accelerators for offloading from server processors data-intensive workloads of a cloud, such as used in collective communication of training data as well as in-network computing operations for reducing/combining that data, which is critical in training large language models for AI.

131. The '297 Patent solves a technological problem with server systems. For example, the '297 Patent explains that “[n]etworked applications often run on dedicated servers that support an associated ‘state’ context or session-defined application. Servers can run multiple applications, each associated with a specific state running on the server.” '297 Patent, 1:23–26. “Unfortunately, servers can be limited by computational and memory storage costs associated with switching between applications. When multiple applications are constantly required to be available, the overhead associated with storing the session state of each application be result in poor performance due to constant switching between applications.” *Id.*, 1:32–38. “Dividing applications between multiple processor cores” does not solve the problem, “since even advanced processors often only have eight to sixteen cores, while hundreds of application or session states may be required.” *Id.*, 1:38–42.

132. To address this issue, the '297 Patent discloses and claims improved systems and methods with computation modules, or DPUs, also referred to as offload processor modules, that can run such session-defined applications in part or full. *See id.*, 2:10–18. In one embodiment, “[i]n effect, one can reduce problems associated with session limited servers by using the module processor (e.g., an ARM architecture processor) of a XIMM to offload part of the functionality of traditional servers.” *Id.*, 2:48–51.

133. The '297 Patent further explains a number of improvements can be achieved using the claimed switching architecture in one embodiment of a new cloud fabric or switching plane “to ingress and egress packets within a parallel mid-plane formed from XIMMs.” *Id.*, 9:25–29. This new switching plane or cloud fabric enables the acceleration of collective communication and reduction/combining operations critical in training large models for AI: “An additional benefit, among others, with such an architecture is the acceleration of Map-Reduce algorithms by an order of magnitude, making them suitable for business analytics.” *Id.*, 9:53–56.

134. For example, Claim 1 of the '297 Patent is directed to:

1. A system, comprising:

a plurality of first server modules interconnected to one another via a communication network, each first server module including

a first switch,

at least one main processor, and

at least one computation module coupled to the main processor by a bus, each computation module including

a second switch, and

a plurality of computation elements; wherein

the second switches of the first server modules form a switching plane for the ingress and egress of network packets independent of any main processors of the first server modules, and

each computation module is insertable into a physical connector of the first server module.

135. For example, Claim 7 of the '297 Patent is directed to:

7. The system of claim 1, wherein the second switch is a virtual switch comprising computation elements on the computation module.

136. NVIDIA is not licensed to the '297 Patent.

137. Microsoft is not licensed to the '297 Patent.

**(2) The '161 Patent – DPU Cloud Network Fabric**

138. U.S. Patent No. 9,378,161 (“the '161 Patent”) is entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors” for forming a new switching plane or cloud fabric that can overcome the speed limitations of the existing cloud network and server systems. The '161 Patent duly and legally issued on June 28, 2016, from U.S. Patent Application No. 13/931,903 (“the '903 Application”), filed on June 29, 2013.

139. The '903 Application was a substitute for U.S. Patent Application No. 61/753,892, filed on January 17, 2013. In addition, the '161 Patent claims priority from U.S. Provisional Application Nos. 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The '161 Patent is entitled to the benefit of these earlier filed applications.

140. Xockets is the current owner of all rights, title, and interest in and to the '161 Patent, including the right to sue for past damages.

141. A true and correct copy of the '161 Patent is attached hereto as **Exhibit 5** and is incorporated by reference herein.

142. The '161 Patent relates to improved systems, hardware, and methods for cloud data centers by creating a rack level server system having offload processor modules, or DPUs, connected together in a novel switching architecture to form a new switching plane or cloud fabric. In particular, the '161 patent relates to rack level or cluster level server systems that include a plurality of servers mountable in a rack and a top of rack (TOR) unit having connections to each of the servers and the existing cloud network. A plurality of offload processor modules are disclosed for offloading data-intensive workloads from server processors of the rack-level server system. Each offload processor module includes multiple offload processors that function as programmable hardware accelerators and at least one input-output (IO) port. The offload processor

modules are connected directly to each other through their respective IO ports and to memory of each server to form a new cloud fabric for cloud offload from server processors, that can bypass the limitations associated with collective communication over the existing cloud network and server systems including conventional Top-Of-Rack switches prone to congestion.

143. The '161 Patent solves a technological problem with server systems., including systems used in data centers and for data processing. For example, the '161 Patent explains that “[e]fficient managing of network packet flow and processing is critical for high performance,” and “[s]ubstantial improvements in network service would be made possible by systems that can flexibly process a data flow, recognize or characterize patterns in the data flow, and improve routing and processing decisions for the data.” ’161 Patent, 1:26–35. “Unfortunately, the tree-like server connection topology often used in conventional data centers can be prone to traffic slowdowns and computational bottlenecks.” *Id.*, 1:36–38. “Typically, all the servers in such data centers communication with each other through higher level Ethernet-type switches, such as Top-Of-Rack (TOR) switches.” *Id.*, 1:38–40. However, as the '161 Patent explains, “[f]low of all the traffic through such TOR switches leads to congestion results in increased network latency, particularly during the periods of high usage.” *Id.*, 1:41–43.

144. Improving upon the prior art, the '161 Patent discloses and claims systems, hardware, and methods relating to rack server systems having DPUs connected together in a novel switching architecture to form a new switching plane or cloud fabric. For example, “to prevent data transfer bottlenecks through TOR switches, and/or to improve[] system performance,” the '161 Patent discloses and claims systems in which “direct inter-rack and/or intra-rack communication can be enabled by offload processor modules included in the servers,” bypassing the limitations associated with the existing cloud network and server systems. *Id.*, 4:19–23. “[S]uch

data communication via offload processor modules can require less time and/or less processing power as compared to TOR switching via aggregation layer transfers. Accordingly, such data transfers can be executed in a more efficient manner than conventional systems.” *Id.*, 4:29–34. In addition, “[a]dvantageously, inter/intra-rack communications via offload server modules can also reduce the need for additional TOR switches and can be included to increase bandwidth and introduce redundancy, particularly since TOR switches may have to be periodically replaced to handle higher network speeds.” *Id.*, 4:35–40.

145. For example, Claim 1 of the ’161 Patent is directed to:

1. A rack server system for a packet processing, comprising:

a plurality of servers mountable in a rack;

a top of rack (TOR) unit having connections to each of the servers;

a plurality of offload processor modules, each offload processor module having at least one input-output (IO) port and multiple offload processors, including at least a first offload processor module connected directly to a second offload processor module through their respective IO ports, the offload processor modules are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module; and

a memory controller configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

146. NVIDIA is not licensed to the ’161 Patent.

147. Microsoft is not licensed to the ’161 Patent.

### **(3) The ’092 Patent – DPU In-Network Computing**

148. U.S. Patent No. 10,212,092 (“the ’092 Patent”) is entitled “Architectures and Methods for Processing Data in Parallel Using Offload Processing Modules Insertable Into Servers” for in-network computing operations on data-intensive workloads in the new switching

plane or new cloud fabric of the '297 and '161 Patents. The '092 Patent duly and legally issued on February 19, 2019, from U.S. Patent Application No. 15/396,330, filed on December 30, 2016.

149. The '092 Patent is a continuation of U.S. Patent Application No. 13,900,318, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012; and U.S. Provision Application Nos. 61/753,892; 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The '092 Patent is entitled to the benefit of these earlier filed applications.

150. Xockets is the current owner of all rights, title, and interest in and to the '092 Patent, including the right to sue for past damages.

151. A true and correct copy of the '092 Patent is attached hereto as **Exhibit 6** and is incorporated by reference herein.

152. The '092 Patent relates to a cloud distributed computing architecture for executing at least first and second computing operations in parallel for processing data-intensive workloads of server processors, including CPUs, GPUs, or hybrids of these server processors. The distributed computing architecture includes a plurality of servers, each server having an offload processing module, or DPU, and a virtual switch along with computation elements that function as hardware accelerators that can be formed into programmable logic pipelines by the virtual switch to offload data-intensive workloads from server processors for executing the second computing operations. These second computing operations can include, for example, collective communication of training results in training large language models for AI as well as in-network reduction/combining of the training results in the new switching plane or cloud fabric that is formed with the offload processing modules, or DPUs. The reduction/combining of training results structures the data for use by the GPUs in further training of large language models. These second computing operations

are performed on first processed data in the claimed invention that can include, for example, the training results that are generated by first computing operations on the GPUs in training large language models for AI. As described earlier, the '092 patent further describes that the offload processing modules can form a new switching plane or cloud fabric using their virtual switches for exchanging the training results between the offload processing modules and performing the second computing operations on the plurality of the offload processing modules in parallel for in-network computing of data-intensive workloads.

153. The '092 Patent solves a technological problem with server architectures for processing data. For example, the '092 Patent explains that “[e]nterprises store and process their large amounts of data in a variety of ways.” '092 Patent, 1:31–32. One manner involves structured data (e.g., data stored in relational databases); “[h]owever, it is estimated that such formatted structured data represents only a tiny fraction of an enterprise’s stored data.” *Id.*, 1:32–41. “Organizations are becoming increasingly aware that substantial information and knowledge resides in unstructured data (i.e., ‘Big Data’) repositories.” *Id.*, 1:41–44. But as the '092 Patent explains, “conventional platforms that are currently being used to handle structured and unstructured data can substantially differ in their architecture.” *Id.*, 1:47–49. Accordingly, the '092 Patent recognized that “[a]n architecture that supports both structured and unstructured queries can better handle current and emerging Big Data applications.” *Id.*, 1:55–57.

154. To address this issue, the '092 Patent discloses improved server architectures for processing data-intensive computing operations in parallel by utilizing offload processing modules, or DPUs, which the '092 Patent in one embodiment refers to as Xocket In-Line Memory Modules (XIMMs), to execute in-network computing operations in a new switching plane or cloud fabric. As the '092 Patent explains, “[d]ata processing and analytics for enterprise server or cloud

based data systems, including both structured or unstructured data, can be efficiently implemented on offload processing modules.” *Id.*, 2:51–54.

155. Using one or more offload processing modules, or DPUs, “it is possible to execute lightweight data processing tasks without intervention from a main server processor.” *Id.*, 2:56–61. In addition, “XIMM modules have high efficiency context switching, high parallelism, and can efficiently process large data sets. Such systems as a whole are able to handle large database searching at a very low power when compared to traditional high power ‘brawny’ server cores.” *Id.*, 2:61–66. Furthermore, “[a]dvantageously, by accelerating implementation of MapReduce or similar algorithms on unstructured data . . . , a XIMM based architecture capable of partitioning tasks is able to greatly improve data analytic performance.” *Id.*, 2:66–3:4.

156. For example, Claim 1 of the ’092 Patent is directed to:

1. A distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data, the architecture comprising:

a plurality of servers, including first servers that each include

at least one central processing unit (CPU), and

at least one offload processing module coupled to the at least one CPU by a bus, each offload processing module including a plurality of computation elements, the computation elements configured to

operate as a virtual switch, and

execute the second computing operations on first processed data to generate second processed data; wherein

the virtual switches form a switch fabric for exchanging data between the offload processing modules,

the first computing operations generate the first processed data and are not executed by the offload processing modules, and

the second computing operations are executed on a plurality of the offload processing modules in parallel.



157. NVIDIA is not licensed to the '092 Patent.

158. Microsoft is not licensed to the '092 Patent.

**(4) The '640 Patent – DPU In-Network Computing**

159. U.S. Patent No. 9,436,640 (“the '640 Patent”) is entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors” for in-network computing operations, including in sorting, organizing, and reducing/combining data-intensive workloads in the new switching plane or new cloud fabric of the '297 and '161 Patents, also known to those of skill in the art as map/reduce operations. The '640 Patent duly and legally issued on September 6, 2016, from U.S. Patent Application No. 13/931,910, filed on June 29, 2013.

160. The '640 Patent claims priority from U.S. Provisional Application Nos. 61/753,892; 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The '640 Patent is entitled to the benefit of these earlier filed applications.

161. Xockets is the current owner of all rights, title, and interest in and to the '640 Patent, including the right to sue for past damages.

162. A true and correct copy of the '640 Patent is attached hereto as **Exhibit 7** and is incorporated by reference herein.

163. The '640 Patent generally relates to systems, hardware, and methods for cloud data centers to create a rack server system with offload processor modules, or DPUs, for in-network reduction/combining of data-intensive workloads using map/reduce data processing, which is critical in training large language models for AI. The rack server system includes a plurality of servers arranged in a rack, and a plurality of offload processor modules, or DPUs, supported on the servers. Each offload processor module has multiple offload processors that function as programmable hardware accelerators and an input-output (IO) port. The offload processor modules

are connected directly to each other through their respective IO ports to form a midplane switch fabric for cloud offload of data-intensive workloads from server processors that can overcome the limitations associated with the existing cloud network and server systems. The offload processor modules are configured to execute map and reduce steps, thereby accelerating map/reduce data processing including collective communication of training results in training large language models for AI as well as in-network reduction/combining of the training results.

164. Like the '161 Patent, the '640 Patent solves a technological problem with server systems, including systems used in data centers and for data processing. For example, the '640 Patent explains that “[e]fficient managing of network packet flow and processing is critical for high performance,” and “[s]ubstantial improvements in network service would be made possible by systems that can flexibly process a data flow, recognize or characterize patterns in the data flow, and improve routing and processing decisions for the data.” '640 Patent, 1:25–34. “Unfortunately, the tree-like server connection topology often used in conventional data centers can be prone to traffic slowdowns and computational bottlenecks.” *Id.*, 1:35–37. “Typically, all the servers in such data centers communication with each other through higher level Ethernet-type switches, such as Top-Of-Rack (TOR) switches.” *Id.*, 1:37–40. However, as the '640 Patent explains, “[f]low of all the traffic through such TOR switches leads to congestion results in increased network latency, particularly during the periods of high usage.” *Id.*, 1:40–43.

165. Improving upon the prior art, the '640 Patent discloses and claims systems, hardware, and methods relating to rack server systems for packet processing. For example, “to prevent data transfer bottlenecks through TOR switches, and/or to improve[] system performance,” the '640 Patent discloses systems in which “direct inter-rack and/or intra-rack communication can be enabled by offload processor modules included in the servers,” bypassing the existing cloud

network. *Id.*, 4:20–24. “[S]uch data communication via offload processor modules can require less time and/or less processing power as compared to TOR switching via aggregation layer transfers. Accordingly, such data transfers can be executed in a more efficient manner than conventional systems.” *Id.*, 4:31–35. In addition, “[a]dvantageously, inter/intra-rack communications via offload server modules can also reduce the need for additional TOR switches and can be included to increase bandwidth and introduce redundancy, particularly since TOR switches may have to be periodically replaced to handle higher network speeds.” *Id.*, 4:36–41.

166. As further explained in the ’640 Patent, “[s]ervers equipped with offload processor modules, such as described herein and equivalent, can bypass a TOR switch through intelligent virtual switching of the offload processor modules associated with each server” and provide in-network computing operations for map/reduce data processing. *Id.*, 18:38–42.

167. For example, Claim 9 of the ’640 Patent is directed to:

9. A rack server system for a map/reduce data processing, comprising:

a plurality of servers arranged in a rack,

a plurality of offload processor modules supported on at least two of the servers, each offload processor module having an input-output (IO) port and multiple offload processors, a first offload processor module configured to execute map steps of the map/reduce data processing, and being connected directly to a second offload processor through their respective IO ports to define a midplane switch, and

a top of rack (TOR) unit connected to each of the servers that does not transfer map/reduce data, wherein

a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module.

168. NVIDIA is not licensed to the ’640 Patent.

169. Microsoft is not licensed to the ’640 Patent.

## II. BACKGROUND ON XOCKETS' CLOUD COMPUTING INVENTIONS

170. The advent of cloud computing has radically changed the computing industry. Instead of individual businesses running a relatively small number of general-purpose server platforms in-house, a relatively few number of specialists have arisen to run massive warehouses of computers. These specialists—cloud operators—could reap efficiencies of scale that individual owners could never reach. The cloud computing model has had profound economic, competitive, and technological implications for the industry—and society as a whole.

171. At first, these cloud operators built systems that looked like a traditional data center, just scaled up. Their systems revolved around server processors (Central Processing Units (CPUs), Graphical Processing Units (GPUs), and hybrids of these server processors). To meet the increased data demands of the growing cloud market, operators of these server processor-based cloud systems expected to rely on the performance increase of each new generation of server processors. These types of processors are optimized for complex programs that make lots of interdependent decisions and perform complicated math. Internally, they spend a tremendous amount of energy in identifying shortcuts through those interdependent decisions and include massive and power-hungry structures for number crunching.

172. And in fact, for the bulk of the history of the processor, the industry could count on steady performance increases in each new generation, resulting from the predictable density improvement of transistors, known as Moore's Law. For a given power budget, performance improvement in processors was essentially guaranteed. Designers could add complexity to software and count on transistor performance to accelerate server processors to the point where they could handle it.

173. Dr. Dalal foresaw that Moore's Law would end, and its death would come at a most inopportune time, as he also anticipated that the data demands of cloud computing would

exponentially increase. He turned out to be right on both scores. Transistor performance scaling through increased transistor density has in fact slowed and, for some metrics, has essentially stopped. Yet the amount of data traffic in the cloud has increased exponentially. We now live in the “Zettabyte Era.” A zettabyte is the current largest digital unit of measurement. A zetta-stack of dollar bills would reach from the earth to the sun (93 million miles away) and back—700,000 times. According to Google’s former CEO Eric Schmidt, from the beginning of humanity to the year 2003, an estimated 0.5% of a zettabyte was created.<sup>31</sup> In 2012, the year that Xockets filed its first provisional patent application, the amount of all digital data in the world first exceeded a zettabyte. Today, the volume of cloud traffic alone is estimated to be 50 zettabytes a year and growing. This is an almost inconceivable increase in data.

174. Anticipating the death of Moore’s Law at a time when cloud data demands would exponentially increase, Dr. Dalal recognized that server processors would become a bottleneck, hitting walls of efficiency and performance. The conventional wisdom of adding more, ever-faster server processors would not solve the problems that the cloud computing platforms of the future would face (and help cause). These were the wrong kind of processors for processing the data-intensive workloads required in cloud distributed computing. First, adding more and more server processors was a significant cost. Second, it did not address the root of the problem, which was that server processors were not designed to efficiently handle the data intensive infrastructure services that are required when massive amounts of data and processing are being pushed to the cloud. No matter how many server processors were added, bottlenecks from the data intensive infrastructure services would continue to arise, interrupting the server processors from running

---

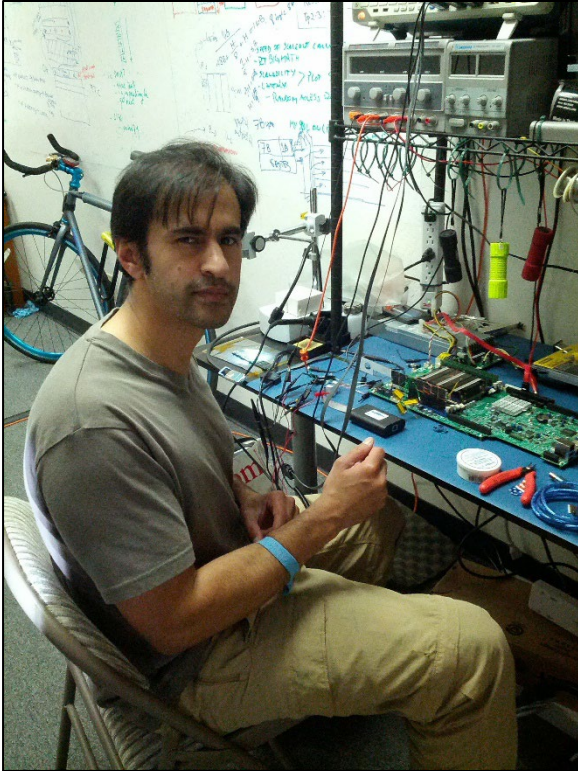
<sup>31</sup> Google Chief Eric Schmidt Keynote Speech at Guardian Activate 2010, Part 2, <https://youtu.be/jcBPgEGA7Yk?si=WpIFgbu3Kwjr39Hw&t=227> (3:47–4:05).

what they were designed to run: revenue-producing cloud applications, including training large language models for AI.

175. Dr. Dalal founded Xockets in 2012. He raised seed funding and then hired a team of network infrastructure engineers to help him develop his technology.

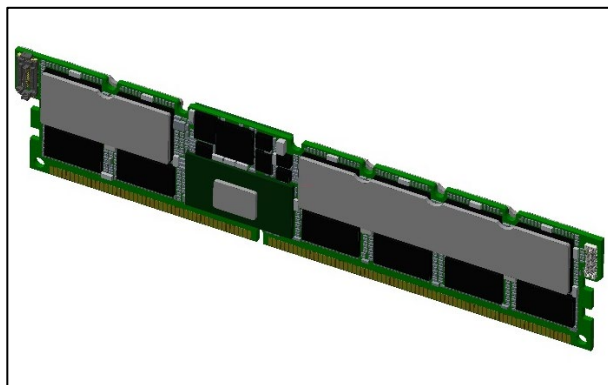
176. In May 2012, the first Xockets provisional patent application was filed. Additional provisional patent applications were filed in January 2013 and March 2014. These provisional patent applications disclosed, among other things, Xockets' DPU-based cloud architecture which *offloads* from CPUs, GPUs, and/or other host processors in servers *and accelerates* the data plane and/or control plane of cloud infrastructure services *independent* of server processors. For example, in Xockets' patented DPU-based cloud architecture, packet processing operations of key infrastructure services in cloud distributed computing may be offloaded from server processors to DPUs, including cloud-specific security, networking, and storage infrastructure services. As another example, brokering and accelerating communications between server processors (such as GPUs) in training large language models for implementing machine learning/artificial intelligence in cloud applications, referred to as cloud "ML/AI collective communications," and related computational operations, for sorting, organizing, and reducing/combining training results, may also be offloaded from server processors to DPUs.

177. Dr. Dalal and his team at Xockets proceeded to design, develop, and build the world's first DPU for cloud offload processing.



178. In late September of 2015, Xockets demonstrated its DPU computing and switching architectures implemented in its StreamSwitch product at Strata, the industry's premier big data and network technology conference.

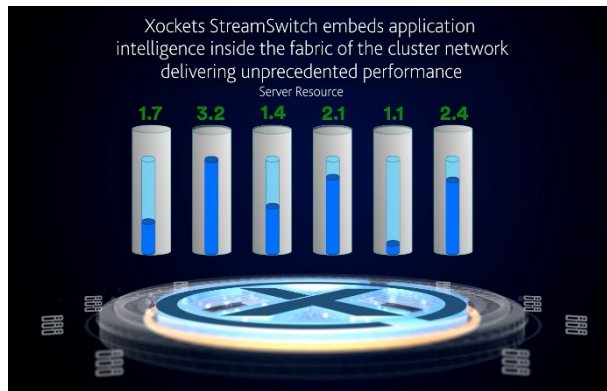
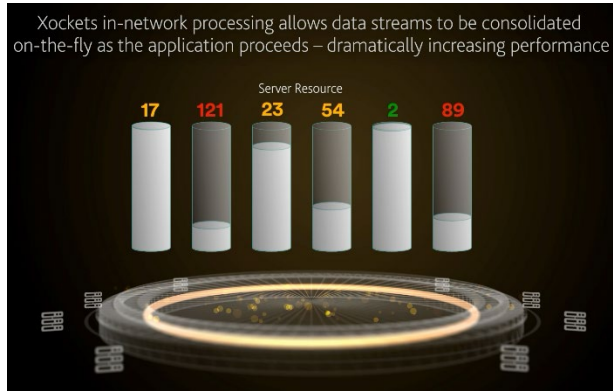




179. As Xockets recognized and showed, distributed computing in a cloud requires computing operations on billions of packets per second flowing between servers, including for cloud-specific security, networking, and storage, and the brokering of collective communications across server processors and in-network reduction of the data for further processing. Xockets'

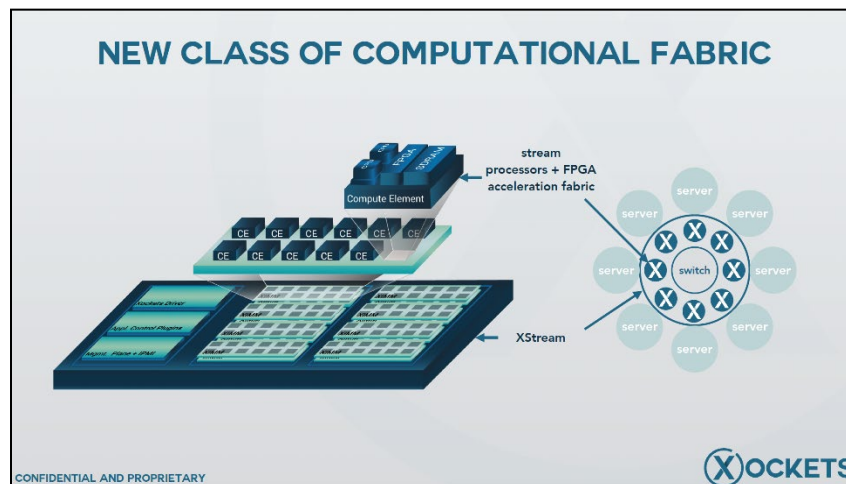


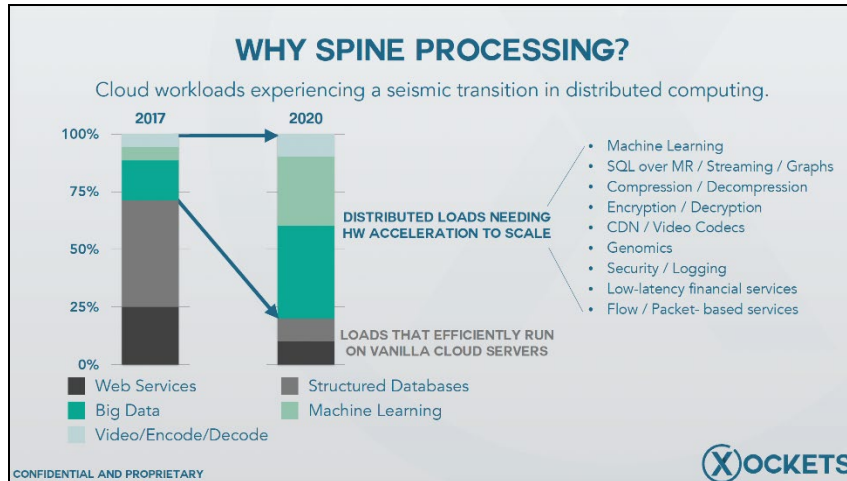
presentations demonstrated the need for programmable hardware acceleration to scale distributed computing in modern cloud workloads, including cloud-specific “security,” “flow/packet-based services,” “big data,” “encryption/decryption,” and “machine learning” workloads for AI.



**The result:**

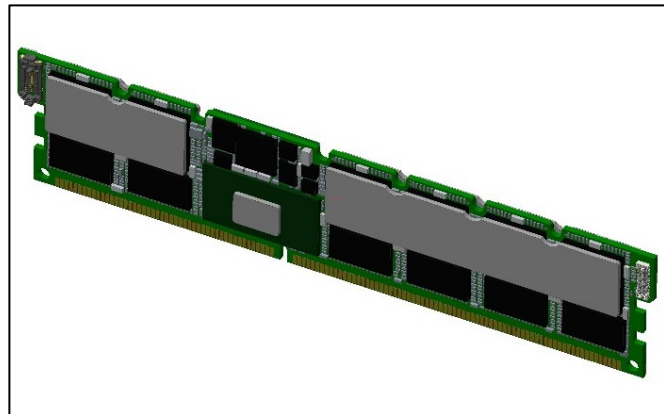
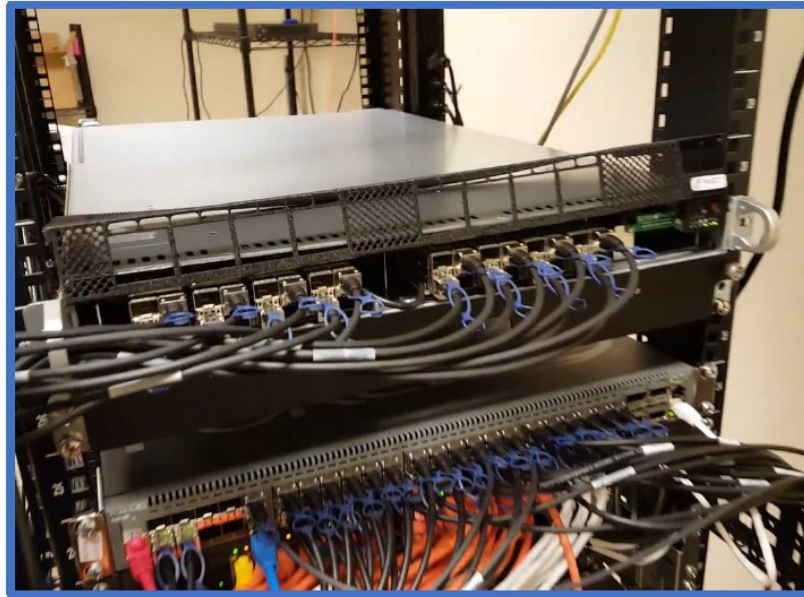
- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure





180. The StreamSwitch included 14 Xockets DPUs, one for each of 14 servers in a standard server rack. The DPUs were branded “Xocket In-line Memory Modules (XIMMs)” in this top of rack embodiment and provided “bump-in-the-wire acceleration” at “line rate” (i.e., network speed, then 10 Gbps) at the boundary between the network and each server processor and were connected at the network server interface through to the server’s system bus.

181. This DPU embodiment included a *Xockets virtual switch computing architecture* for identifying and classifying packet flows and connecting together programmable logic pipelines of hardware accelerators for performing computational operations on the packet data for bump-in-the-wire offload and hardware acceleration of the cloud data plane and/or control plane independent of server processors. The photographs below show a Xockets StreamSwitch at Top of Rack, and the inside of a StreamSwitch, depicting the Xockets DPUs:



182. The Xockets team achieved up to ***1000x acceleration*** in the performance of Hadoop big data applications running on the StreamSwitch compared to host CPUs performing the same cloud-specific data plane infrastructure services. This performance benefit would result in (1) substantially higher revenues in a cloud data center as a result of freeing up server processors

for running more revenue-producing customer applications and services and by enabling accelerated computing performance in a cloud; and (2) substantially lower Total Cost of Ownership (TCO) for servers deployed in a cloud data center, resulting from a reduction of both capital expenses (fewer servers/CPUs for the same performance) and operating expenses (lower energy costs for the same performance).

183. After years of heading in the wrong direction, and following the publication of Xockets' patent applications and its public demonstrations of its pioneering new DPU-based architecture, NVIDIA abandoned conventional server processor-centric approaches and embraced implementing Xockets' patented architecture by utilizing DPUs for cloud offload processing. Xockets' patented DPU-based cloud architecture is now [REDACTED]

### III. NVIDIA'S USE OF XOCKETS' PATENTED TECHNOLOGY

184. NVIDIA began its business in the graphics market. However, users soon realized that NVIDIA's graphics processors could be used as very high-performance server processors to perform massive technical number crunching. This led NVIDIA to begin selling more expensive processors into more lucrative markets such as cryptocurrency.

185. However, NVIDIA continued to operate as a component vendor for many years. In 2012, NVIDIA graphics processors were added to the Department of Energy's Titan supercomputer. But the complete system was designed and built by Cray, using AMD microprocessors and Cray-designed networking. This system was followed by two more supercomputers, Summit and Sierra, which were designed by IBM using NVIDIA GPUs and Mellanox-designed networking.

186. NVIDIA eventually realized that it needed to move from being a component vendor to being a system designer and provider. NVIDIA explained: "While computing demand is

surging, CPU performance advances are slowing as Moore’s law has ended. This has led to the adoption of accelerated computing with NVIDIA GPUs and Mellanox’s intelligent networking solutions. Datacenters in the future will be architected as giant compute engines with tens of thousands of compute nodes, designed holistically with their interconnects for optimal performance.”<sup>32</sup> This solution NVIDIA hit upon was to take Xockets’ patented technology.

187. NVIDIA uses Xockets’ New Cloud Processor and New Cloud Fabric technology.

**A. NVIDIA’S INFRINGEMENT OF THE NEW CLOUD PROCESSOR PATENTS**

188. NVIDIA infringes the ’209, ’924, and ’350 Patents in at least its DPU-enabled cloud computing systems, for example, NVIDIA’s Hopper and Blackwell GPU-enabled server computer systems available in DGX, HGX, MGX, and other configurations utilizing at least NVIDIA BlueField DPUs and/or ConnectX DPUs (ConnectX-5 and later versions) (hereinafter, the “NVIDIA Accused Products for New Cloud Processor Patents”). The ’209 Patent describes Xockets’ DPU computing architecture wherein DPUs (“computation modules”) with a “virtual switch” computing architecture perform data-intensive computing operations “independent of server processors” in cloud data centers. The ’924 Patent describes Xockets’ DPU computing architecture wherein DPUs “independent of server processors” form “network overlays” in cloud data centers. Together, the ’209 and ’924 Patents describe and claim offloading, accelerating, and isolating cloud VPN and other network security services using Xockets’ DPU computing architecture. The ’350 Patent describes “hardware acceleration modules,” DPUs, that include “scheduler circuit” technology for implementing stream processing in general-purpose processors (e.g., ARM cores) to function as general-purpose hardware accelerators at the speed of the network

---

<sup>32</sup> <https://nvidianews.nvidia.com/news/nvidia-to-acquire-mellanox-for-6-9-billion>.

or line rate to enable full programmability and versatility for in-network computing operations in cloud data centers.

189. Xockets' Patents, including its New Cloud Processor Patents that protect its DPU computing architecture, provide the versatility needed in offloading, accelerating, and isolating from cloud server processors the data-intensive computing tasks required in making distributed computing and AI possible in data centers.

190. NVIDIA not only infringes the '209, '924, and '350 Patents, it publicly touts its use of the claimed technology and the performance benefits it provides.

191. NVIDIA's CEO Huang has consistently touted NVIDIA's use of Xockets' patented technology.

192. In April 2020, Huang appeared on CNBC to "[take] a victory lap after his company completed its long-awaited merger of chip producer Mellanox Technologies"—a \$7 billion deal<sup>33</sup>:

“Man I’ve been dreaming about this. You know the most important computer today is the data center, it is the epicenter of the computer industry. And *the most important applications that run in the data center today are AI applications and Big Data analytics applications. Doing computation on artificial intelligence . . . and moving huge amounts of data around is what drives the data center architectures today.* And so we are combining the leaders of AI computing and high speed networking and data processing into one company. This is really quite extraordinary.”<sup>34</sup>

193. Describing NVIDIA's adoption of the DPU computing architecture protected by Xockets' patents, NVIDIA stated:

---

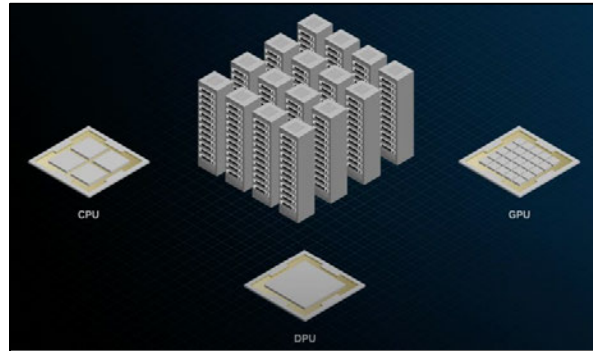
<sup>33</sup> <https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html>.

<sup>34</sup> <https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html>.



“Data centers are evolving and expanding to include another pillar along the CPU and GPU. ***The new pillar is the Data Processing Unit or DPU.***

The NVIDIA BlueField DPU is designed to ***offload, accelerate, and isolate*** infrastructure workloads and bring efficiency and security to software defined [data-intensive] workloads such as networking security and storage while freeing CPU resource by up to 30%.”<sup>35</sup>



194. NVIDIA, including its CEO Huang, agrees that DPUs use a fundamental new computing architecture:

A DPU is going to be programmable . . . and it’s going to ***offload the movement of data into the granular processing of the data as it’s being transmitted and keep it from ever bothering the CPUs and GPUs*** and avoid redundant copies of data. ***That’s the architecture of the future. And that’s the reason why we’re so excited about Mellanox.***<sup>36</sup>

195. In fact, according to Huang, the architecture is revolutionary:

***I really do think that when you offload [to] the data processing on the SmartNIC [DPU], when you’re able to disaggregate the converged server, when you can put accelerators anywhere in datacenter and then can compose and reconfigure that datacenter for this specific workload – that’s a revolution.***<sup>37</sup>

---

<sup>35</sup> NVIDIA DOCA Software Framework, <https://www.youtube.com/watch?v=htR19rdBicA&t=3s> (0:03–0:30) (illustrating the operation of NVIDIA’s BlueField and ConnectX DPUs).

<sup>36</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

<sup>37</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

196. NVIDIA also lauds this technology as pioneering—a “reinvention from the ground up”—even though Xockets invented, patented, and disclosed it to the world years before NVIDIA first implemented it:

“There are two fundamental transitions happening in the computer industry today.

The first trend is because CPU scaling [Moore’s Law] has ended. . . . [W]e need a new computing approach and accelerated computing is the path forward.

It happened at exactly the time when a new way of doing software was discovered, deep learning [ML/AI], these two events came together and it’s driving computing today: *Accelerated computing and generative AI. . . . This way of doing computation is a reinvention from the ground up.*”<sup>38</sup>



197. NVIDIA has also repeatedly recognized the importance of DPUs for cloud offload processing, and described the benefits achieved by implementing Xockets’ patented inventions.

198. For example, introducing the NVIDIA BlueField-2 DPU at NVIDIA’s GTC 2020 Keynote on October 5, 2020, Huang said:

---

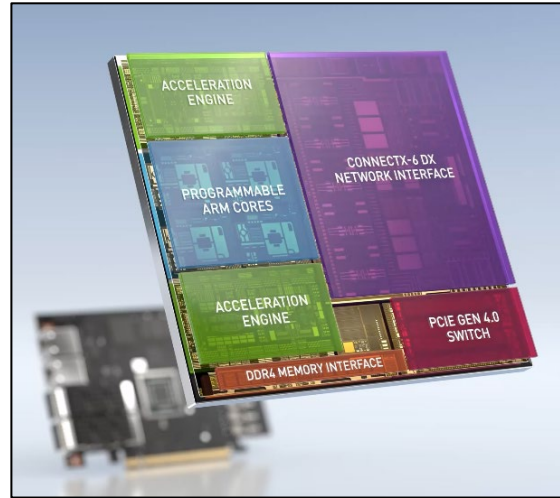
<sup>38</sup> NVIDIA Keynote at COMPUTEX 2023, <https://www.youtube.com/watch?v=i-wpzS9ZsCs&t=874s> (14:34–15:46).



“Today we are announcing the BlueField-2 DPU.

It is a *programmable processor with accelerators and engines for at-line-speed processing for networking, storage, and security*. The BlueField DPU is a data center infrastructure on a chip. BlueField-2 has Arm CPUs and a whole host of state-of-the-art accelerators and hardware engines. BlueField-2 does the security processing for private, public, and hybrid clouds. . . .

BlueField-2 is a 7 billion transistor marvel. A programmable data center on-a-chip. One that we intend to support for as long as we shall live.”<sup>39</sup>



199. NVIDIA further stated that its DPUs “enable[] breakthrough networking, storage and security performance”—as claimed in the New Cloud Processor Patents.<sup>40</sup>

200. In addition, Huang directly explained how DPUs enable breakthrough performance in clouds.<sup>41</sup>

201. And NVIDIA touts the ability to “transform the data center with NVIDIA DPUs” by “offloading, accelerating, and isolating a broad range of advanced networking, storage, and security services.”<sup>42</sup>

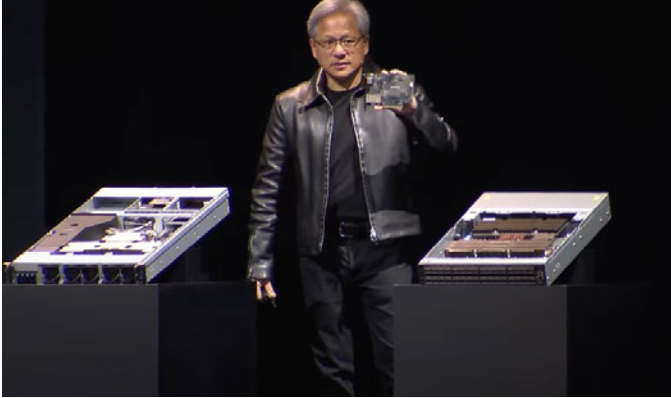
202. NVIDIA further characterized its infringing DPU-enabled architecture—first described and claimed by Xockets—as essential for high performance networks necessary for generative AI:

<sup>39</sup> NVIDIA CEO Jensen Huang’s Keynote at GTC – Part 5 – October 2020, <https://youtu.be/MRdJ78dWjn4?t=138> (2:18–4:15).

<sup>40</sup> <https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center>.

<sup>41</sup> <https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang>.

<sup>42</sup> <https://www.advancedhpc.com/pages/nvidia-bluefield-data-processing-units>.



*“Our new Ethernet system for AI is the Spectrum-4 switch and **the BlueField-3 SmartNIC or DPU** . . . This is what it takes to build a high performance network. And we’re gonna take this capability to the world’s CSPs.*

*The reception has been incredible, and the reason for that is, of course, every CSP, every data center would like to turn every single data center into a generative AI data center.”<sup>43</sup>*

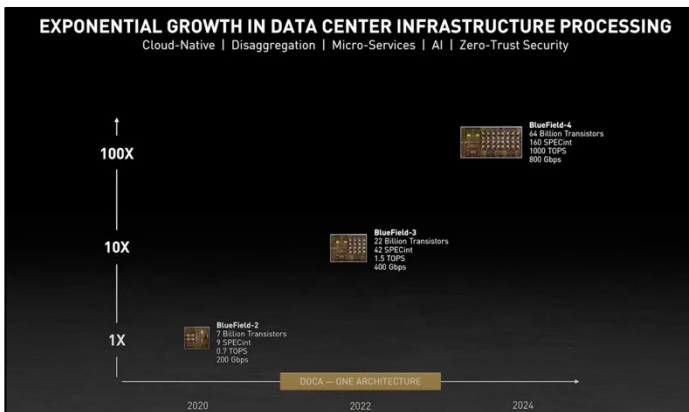
203. As stated on NVIDIA’s website, the BlueField DPU “ignites unprecedented innovation for modern data centers and supercomputing clusters. With its robust compute power and integrated software-defined hardware accelerators for networking, storage, and security, BlueField creates a secure and accelerated infrastructure for any workload in any environment, ushering in a new era of accelerated computing and AI.”<sup>44</sup>

204. Huang explained this further during his keynote address at the 2021 GPU Technology Conference:

---

<sup>43</sup> NVIDIA Keynote at COMPUTEX 2023, <https://www.youtube.com/watch?v=i-wpzS9ZsCs&t=4879s> (1:21:19–1:22:21).

<sup>44</sup> <https://www.nvidia.com/en-in/networking/products/data-processing-unit>.

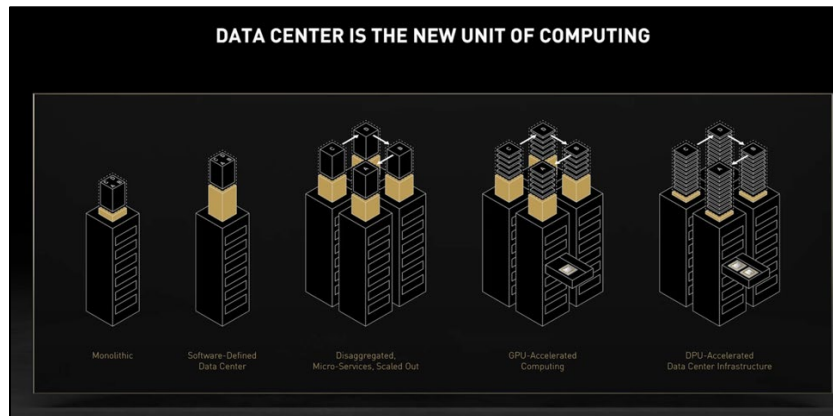


“A simple way to think about this is that *one-third of the roughly 30 million data center servers shipped each year are consumed running the software-defined data center stack.* This workload is increasing much faster than Moore’s law.

We know this because of the amount of data we are producing and moving around. *So, unless we offload and accelerate this workload, data centers will have fewer and fewer CPUs to run applications. The time for BlueField has come.*”<sup>45</sup>

205. Huang continued:

“*Data center is the new unit of computing. Cloud computing and AI are driving fundamental changes in the architecture of data centers.* Traditionally, enterprise data centers ran monolithic software packages. Virtualization started the trend toward software-defined data centers – allowing applications to move about . . . With virtualization, *the compute, networking, storage, and security functions are emulated on software running on the CPU.* Though easier to manage, *the added CPU load reduced the data center’s capacity to run applications,* which is its primary purpose.



*This illustration shows the added CPU load in the gold-colored part of the stack.* Cloud computing re-architected data centers again,

<sup>45</sup> The Data Center is the New Unit of Computing (NVIDIA GTC 2021 Keynote Part 3), <https://www.youtube.com/watch?v=rzdBHBx3eJk&t=320s> (5:20–5:47).

now to provision services for billions of consumers. Monolithic applications were disaggregated into smaller microservices that can take advantage of any idle resource. . . . Data center networks became swamped by east-west traffic generated by disaggregated microservices. . . . Then, deep learning emerged. . . . Deep learning is compute-intensive, which drove adoption of GPUs. . . . *Meanwhile, the mountain of infrastructure software continues to grow. . . . The answer is a new type of chip for data center infrastructure processing like NVIDIA's BlueField DPU.*"<sup>46</sup>

206. NVIDIA also conceded that DPUs deliver extraordinary TCO Savings in training large models. As NVIDIA has touted: "*A single BlueField-2 DPU can deliver the same data center services that could consume up to 125 CPU cores.* This frees up valuable CPU cores to run a wide range of other enterprise applications."<sup>47</sup>

207. Elsewhere, NVIDIA has also admitted that "*[o]ne Bluefield-3 DPU delivers the equivalent data center services of up to 300 CPU cores,* freeing up valuable CPU cycles to run business-critical applications."<sup>48</sup>

208. On August 30, 2022, Huang further publicly touted NVIDIA's DPUs, and explained:

"The return on investment — the benefits that DPU-enabled vSphere 8 with NVIDIA Bluefield deliver — will be so fast because it frees up so many resources for computing that the payback is going to be instantaneous . . . . It's going to be a really fantastic return."<sup>49</sup>

---

<sup>46</sup> The Data Center is the New Unit of Computing (NVIDIA GTC 2021 Keynote Part 3), <https://www.youtube.com/watch?v=rzdBHBx3eJk&t=4s> (0:04–1:59).

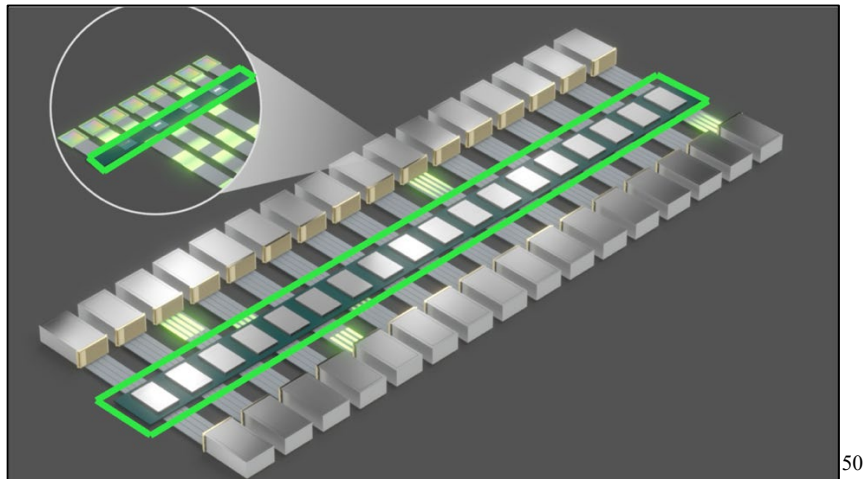
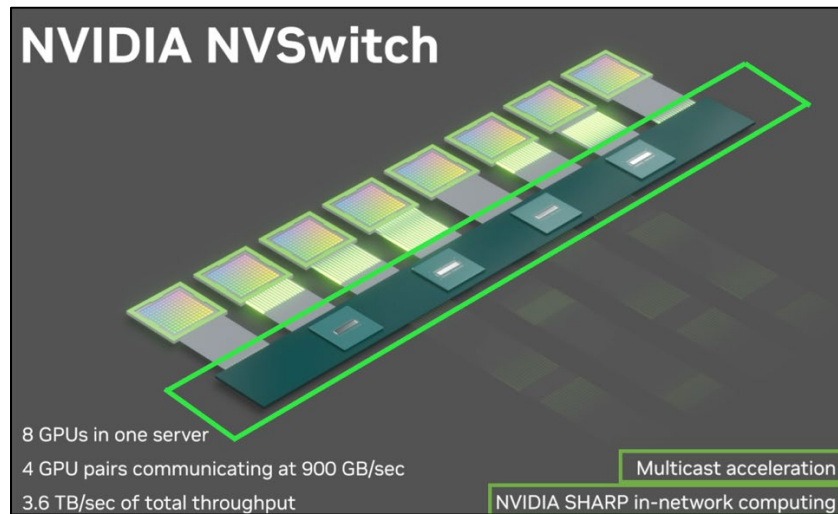
<sup>47</sup> <https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center>.

<sup>48</sup> <https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3>.

<sup>49</sup> <https://blogs.nvidia.com/blog/nvidia-vmware-new-era-enterprise-computing>; see also <https://youtu.be/TuggBO97yYg?si=yxu0LnAN85df-frZ&t=784> (13:04–13:23).

## B. NVIDIA'S INFRINGEMENT OF THE NEW CLOUD FABRIC PATENTS

209. NVIDIA infringes the '297, '161, '092, and '640 Patents in at least its DPU-enabled cloud computing systems, including NVIDIA's Hopper and Blackwell GPU-enabled server computer systems available in DGX, HGX, MGX, and other configurations utilizing at least NVIDIA NVLink Switch DPUs (hereinafter, the "NVIDIA Accused Products for the New Cloud Fabric Patents").



<sup>50</sup> NVIDIA video illustrating the operation of its NVLink Switch fabric, <https://images.nvidia.com/aem-dam/Solutions/gtcs22/nvlink/hpc-video-nvlink-video-2781044.mp4>.

210. As explained above, the '297 Patent describes Xockets' DPU switching architecture wherein computation modules (e.g., NVIDIA's NVLink Switch DPUs), acting "independent of any main processors," form a new "switching plane" or cloud fabric for offloading, accelerating, and isolating data-intensive workloads in clouds, including in training large language models for ML/AI. The '161 Patent describes Xockets' DPU switching architecture employing a plurality of directly connected offload processing modules (e.g., NVIDIA's NVLink Switch DPUs) to form a switching plane or cloud fabric for offloading, accelerating, and isolating data-intensive workloads in clouds that can overcome limitations associated with the existing cloud network and server systems. The '092 Patent describes Xockets' DPU switching architecture wherein offload processing modules (e.g., NVIDIA's NVLink Switch DPUs) operate as virtual switches forming a new switching fabric or cloud fabric for executing in-network computing operations on data-intensive workloads in cloud data centers independent of server processors. The '640 Patent describes Xockets' DPU switching architecture wherein a "plurality of offload processor modules" (e.g., NVIDIA's NVLink Switch DPUs) form a new switching plane or cloud fabric that can overcome limitations associated with the existing cloud network and server systems for implementing in-network computing operations. The '640 Patent further describes "map/reduce data processing" as the in-network computing operations, such as used in reducing/combining training results in training large language models for ML/AI.

211. NVIDIA's infringement of the '297, '161, '092, and '640 Patents in at least its NVLink Switch-based cloud computing systems enables its industry-leading, high-performance training of AI large language models, including by enabling accelerated computing and AI in data centers and turning every data center into an "AI factory."

212. The AI factories NVIDIA sells (comprising GPU servers connected together using Xockets’ DPU inventions, including in DGX, HGX, and MGX configurations) are built to train AI models for producing artificial intelligence for integration into the operations of every business in every industry—and they will usher in a new industrial revolution.

213. NVIDIA holds monopoly power in the market for GPU-enabled artificial intelligence supercomputers that third parties can purchase, holding market share above 90% by unit.

214. NVIDIA not only infringes the ’297, ’161, ’092, and ’640 Patents, it publicly touts its use of the claimed technology and the performance benefits it provides.

215. For example, NVIDIA explains how the NVLink Switch DPUs provide “full connection for unparalleled performance” and are “essential building blocks” for creating “the most powerful AI and HPC [high-performance computing] platform”:

#### **Full Connection for Unparalleled Performance**

*The NVLink Switch is the first rack-level switch chip capable of supporting up to 576 fully connected GPUs in a non-blocking compute fabric.* The NVLink Switch interconnects every GPU pair at an incredible 1,800GB/s. *It supports full all-to-all communication.* The 72 GPUs in GB200 NVL72 can be used as a single high-performance accelerator with up to 1.4 exaFLOPS of AI compute power.

#### **The Most Powerful AI and HPC Platform**

*NVLink and NVLink Switch are essential building blocks of the complete NVIDIA data center solution* that incorporates hardware, networking, software, libraries, and optimized AI models and applications from the NVIDIA AI Enterprise software suite and the NVIDIA NGC™ catalog. *The most powerful end-to-end AI and HPC platform*, it allows researchers to deliver real-world results and deploy solutions into production, *driving unprecedented acceleration at every scale.*<sup>51</sup>

---

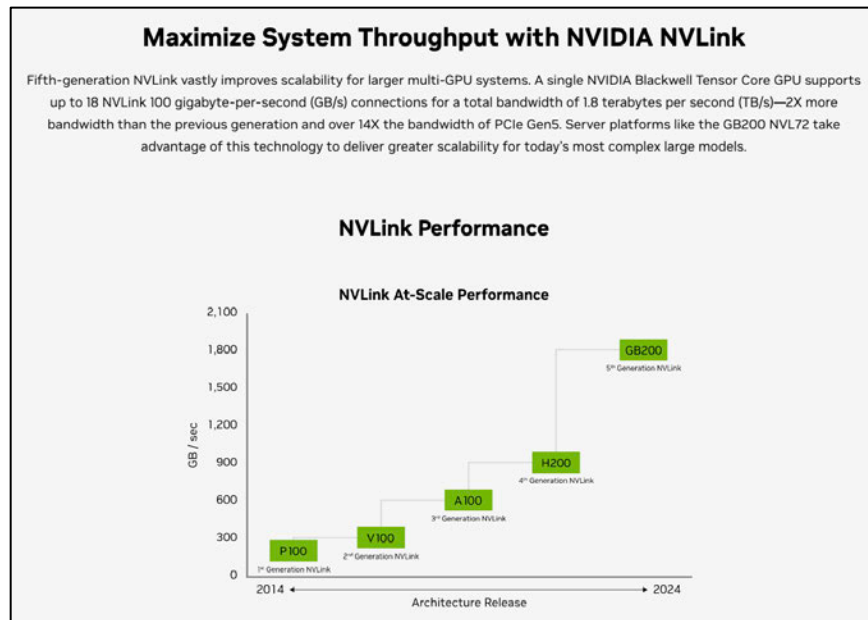
<sup>51</sup> <https://www.nvidia.com/en-us/data-center/nvlink>.



216. NVIDIA further describes the NVLink and NVLink Switch DPUs as “[t]he building blocks of high-speed, multi-GPU communication for feeding large datasets faster into models and rapidly exchanging data between GPUs” and explains how NVLink addresses “a need for faster scale-up interconnects”:

### A Need for Faster Scale-Up Interconnects

Unlocking the full potential of exascale computing and trillion-parameter AI models hinges on swift, seamless communication between every GPU within a server cluster. The fifth generation of NVIDIA NVLink is a scale-up interconnect that *unleashes accelerated performance for trillion- and multi-trillion parameter AI models*.<sup>52</sup>



217. NVIDIA explains how the NVIDIA NVLink and NVLink Switch DPUs “fully connect GPUs” (e.g., clusters of servers in racks known as superpods or supercomputers) in order to “enable high-speed, collective operations” (e.g., map/reduce operations):

<sup>52</sup> <https://www.nvidia.com/en-us/data-center/nvlink>.



### **Fully Connect GPUs With NVIDIA NVLink and NVLink Switches**

NVLink is a 1.8TB/s bidirectional, direct GPU-to-GPU interconnect that scales multi-GPU input and output (IO) within a server. *The NVIDIA NVLink Switch chips connect multiple NVLinks to provide all-to-all GPU communication at full NVLink speed within a single rack and between racks.*

*To enable high-speed, collective operations, each NVLink Switch has engines for NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)<sup>TM</sup> for in-network reductions and multicast acceleration.<sup>53</sup>*

218. In addition, NVIDIA explains how its NVLink Switch DPU is used to “train multi-trillion parameter models”:

### **Train Multi-Trillion Parameter Models With NVLink Switch System**

*With NVLink Switch, NVLink connections can be extended across nodes to create a seamless, high-bandwidth, multi-node GPU cluster—effectively forming a data center-sized GPU.* NVIDIA NVLink Switch enables 130TB/s of GPU bandwidth in one NVL72 for large model parallelism. Multi-server clusters with NVLink scale GPU communications in balance with the increased computing, so NVL72 can support 9X the GPU count than a single eight-GPU system.<sup>54</sup>

219. On information and belief, NVIDIA makes, uses, and sells its NVLink Switch DPUs in forming a new switching plane or cloud fabric for at least its DGX, HGX, and MGX AI supercomputer systems (AI factories).<sup>55</sup>

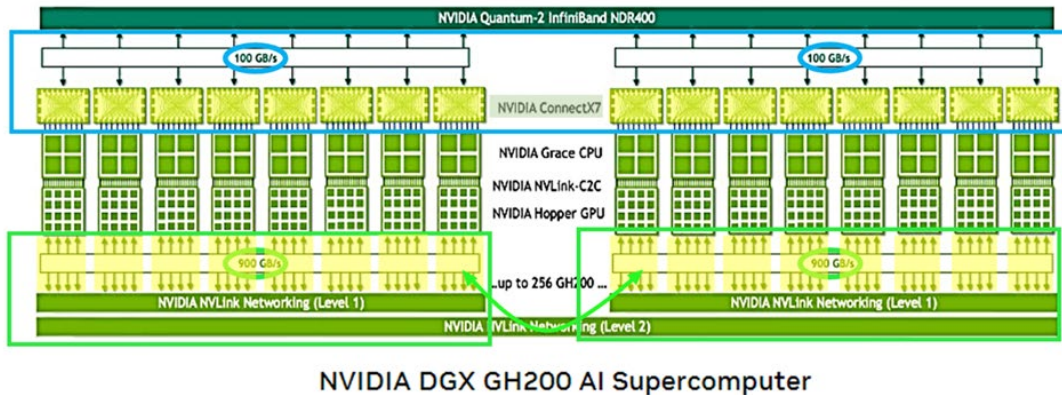
---

<sup>53</sup> <https://www.nvidia.com/en-us/data-center/nvlink>.

<sup>54</sup> <https://www.nvidia.com/en-us/data-center/nvlink>.

<sup>55</sup> See, e.g., <https://www.nvidia.com/en-us/data-center/dgx-platform>; <https://www.nvidia.com/en-us/data-center/hgx>; <https://www.nvidia.com/en-us/data-center/products/mgx>.

Figure 5. NVIDIA NVLink Switch System connects upto 256 NVIDIA Grace Hopper Superchips



NVIDIA DGX GH200 AI Supercomputer

56

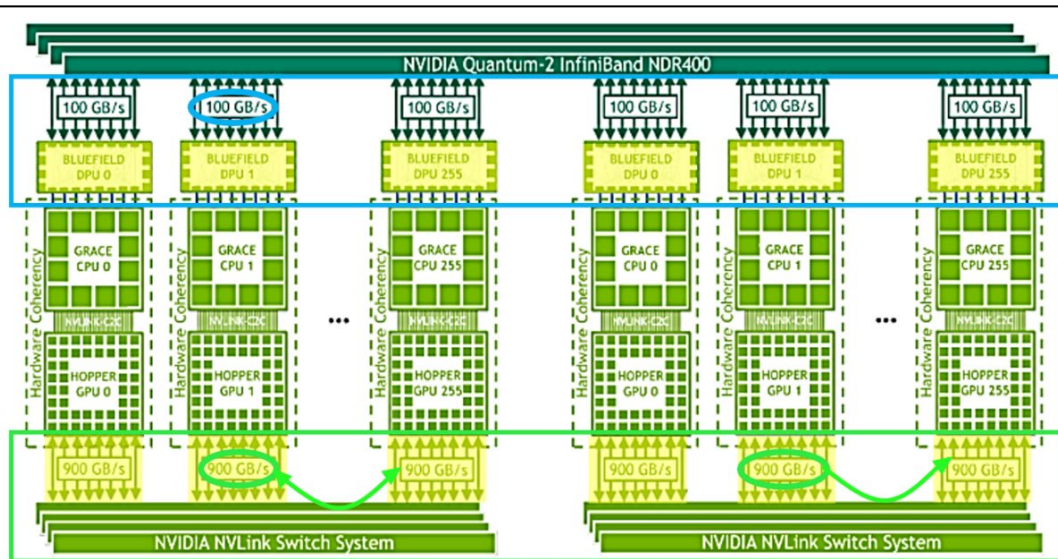


Figure 14. NVIDIA HGX Grace Hopper with NVLink Switch System for strong-scaling giant ML and HPC workloads

57

220. In the third quarter of 2022, NVIDIA released NVLink Switch DPUs for use in forming a new switching plane or cloud fabric for its Hopper H100 GPU servers with hardware accelerators to “upgrad[e] multi-GPU interconnectivity” for AI computing:

<sup>56</sup> <https://resources.nvidia.com/en-us-dgx-gh200/technical-white-paper>.

<sup>57</sup> <https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth>.

*Increasing demands in AI and high-performance computing (HPC) are driving a need for faster, more scalable interconnects with high-speed communication between every GPU.*

*The third-generation NVIDIA NVSwitch is designed to satisfy this communication need. This latest NVSwitch and the H100 Tensor Core GPU use the fourth-generation NVLink, the newest high-speed, point-to-point interconnect by NVIDIA.*

The third-generation NVIDIA NVSwitch is designed to provide connectivity within a node or to GPUs external to the node for the NVLink Switch System. It also incorporates hardware acceleration for collective operations with multicast and NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) in-network reductions.

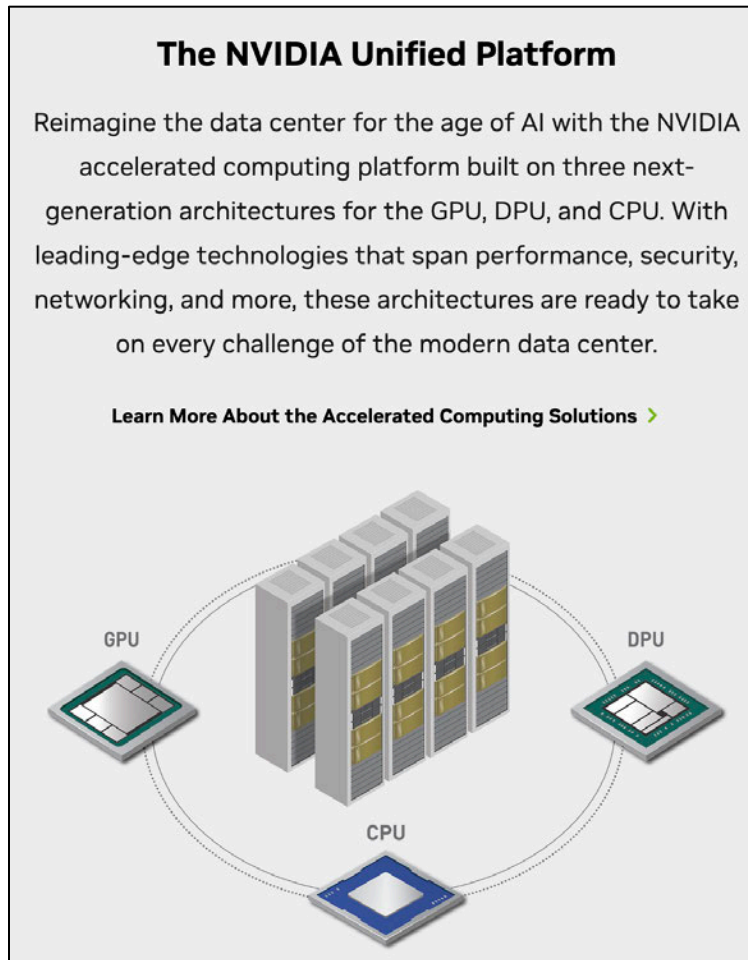
NVIDIA NVSwitch is also a critical enabler of the NVLink Switch networking appliance, which *enables the creation of clusters with up to 256 connected NVIDIA H100 Tensor Core GPUs* and 57.6 TB/s of all-to-all bandwidth. The appliance delivers 9x more bisection bandwidth than was possible with HDR InfiniBand on NVIDIA Ampere Architecture GPUs.<sup>58</sup>

221. On information and belief, at least by March 2024, NVIDIA began advertising “[t]he NVIDIA Unified Platform”—an “accelerated computing platform built on three next-generation architectures for the GPU, DPU, and CPU” that “[r]eimagin[e]d the data center for the age of AI”<sup>59</sup>:

---

<sup>58</sup> Upgrading Multi-GPU Interconnectivity with the Third-Generation NVIDIA NVSwitch, NVIDIA Developer, Aug. 23, 2022, <https://developer.nvidia.com/blog/?p=53977>.

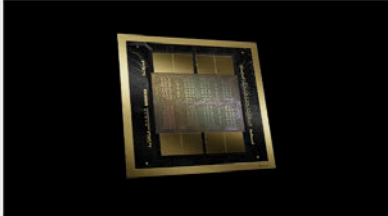
<sup>59</sup> <https://www.nvidia.com/en-us/data-center>; see also <https://web.archive.org/web/20240326035949/https://www.nvidia.com/en-us/data-center>.



222. NVIDIA also highlights below that at least its Blackwell GPU, Grace CPU, and BlueField DPU Architectures will be included in its new Unified AI Platform (also called an AI factory available in DGX, HGX, and MGX configurations that come with a built-in NVLink Switch DPU fabric), stating that “[t]he NVIDIA Blackwell architecture *defines the next chapter in generative AI and accelerated computing with unparalleled performance, efficiency, and scale,*” and “[t]he NVIDIA BlueField architecture is *transforming traditional computing environments into efficient, high- performance, secure, and sustainable data centers for the next wave of applications. By offloading, accelerating, and isolating a broad range of software-*

*defined networking, storage, and security services*, BlueField DPUs create a powerful infrastructure suitable for any workload, in any environment, from cloud to edge”<sup>60</sup>:

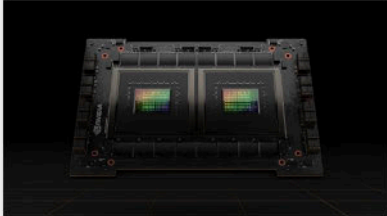
### Architectures for the Modern Data Center



**Blackwell GPU Architecture**

The NVIDIA Blackwell architecture defines the next chapter in generative AI and accelerated computing with unparalleled performance, efficiency, and scale. Blackwell features six transformative technologies that will help unlock breakthroughs in data processing, electronic design automation, computer-aided engineering, and quantum computing.


[Learn More >](#)



**Grace CPU Architecture**

The NVIDIA Grace™ architecture is designed for a new type of emerging data center—AI factories that process and refine mountains of data to produce intelligence. These data centers run a variety of workloads, from AI training and inference, to HPC, data analytics, digital twins, Cloud Graphics and Gaming, and thousands of hyperscale cloud applications.

[Learn More >](#)



**BlueField DPU Architecture**

The NVIDIA BlueField® architecture is transforming traditional computing environments into efficient, high-performance, secure, and sustainable data centers for the next wave of applications. By offloading, accelerating, and isolating a broad range of software-defined networking, storage, and security services, BlueField DPUs create a powerful infrastructure suitable for any workload, in any environment, from cloud to edge.

[Learn More >](#)

223. Further, NVIDIA also markets its infrastructure as “transform[ing] data centers for the era of AI,” and states that “[o]ne-third of the 30 million data center servers shipped every year are consumed by the software-defined data center stack. To support heavy data center workloads, enterprises need to modernize their network infrastructure and evolve to keep up with the exponential demand for data processing.”<sup>61</sup>

224. NVIDIA also describes its BlueField DPU: it “speeds up data center infrastructure workloads, transforming traditional computing environments into efficient, high performance, secure, and sustainable data centers”<sup>62</sup>:

<sup>60</sup> <https://www.nvidia.com/en-us/data-center>.

<sup>61</sup> <https://www.nvidia.com/en-us/networking>.

<sup>62</sup> <https://www.nvidia.com/en-us/networking>.



## BlueField Networking Platform


The NVIDIA BlueField networking platform speeds up data center infrastructure workloads, transforming traditional computing environments into efficient, high-performance, secure, and sustainable data centers, from cloud to edge.

[Learn About BlueField](#)



225. Similarly, NVIDIA markets its DPUs as “transform[ing] the data center with NVIDIA BlueField.” NVIDIA explains the power of DPUs: “The NVIDIA BlueField networking platform *ignites unprecedented innovation for modern data centers* and supercomputing clusters. With its robust compute power and integrated software-defined hardware accelerators for networking, storage, and security, BlueField *creates a secure and accelerated infrastructure for any workload in any environment, ushering in a new era of accelerated computing and AI*”<sup>63</sup>:


### Explore NVIDIA's Portfolio of BlueField Networking Platforms



**NVIDIA BlueField-3 DPU**

The NVIDIA BlueField-3 DPU is a 400 gigabits per second (Gb/s) infrastructure compute platform with line-rate processing of software-defined networking, storage, and cybersecurity. BlueField-3 combines powerful computing, high-speed networking, and extensive programmability to deliver software-defined, hardware-accelerated solutions for the most demanding workloads. From accelerated AI to hybrid cloud, high-performance computing to 5G wireless networks, BlueField-3 redefines the art of the possible.


[Explore BlueField-3 DPUs >](#)



**NVIDIA BlueField-3 SuperNIC**

The BlueField-3 SuperNIC is an advanced network accelerator, purpose-built for supercharging hyperscale AI workloads. Designed for network-intensive, massively parallel computing, the BlueField-3 SuperNIC provides up to 400Gb/s of remote direct-memory access (RDMA) over Converged Ethernet (RoCE) network connectivity between GPU servers, optimizing peak AI workload efficiency. Creating a new era of AI cloud computing, the BlueField-3 SuperNIC enables secure, multi-tenant data center environments with deterministic and isolated performance between jobs and tenants.

[Explore BlueField-3 SuperNICs >](#)



**NVIDIA BlueField-2 DPU**

The NVIDIA BlueField-2 DPU provides innovative acceleration, security, and efficiency in every host. BlueField-2 combines the power of the NVIDIA ConnectX®-6 Dx with programmable Arm® cores and hardware offloads for software-defined networking, storage, security, and management workloads. BlueField-2 delivers superior performance, security, and reduced total cost of ownership for cloud computing platforms, enabling organizations to efficiently build and operate virtualized, containerized, and bare-metal infrastructures at massive scale.

[Explore BlueField-2 DPUs >](#)

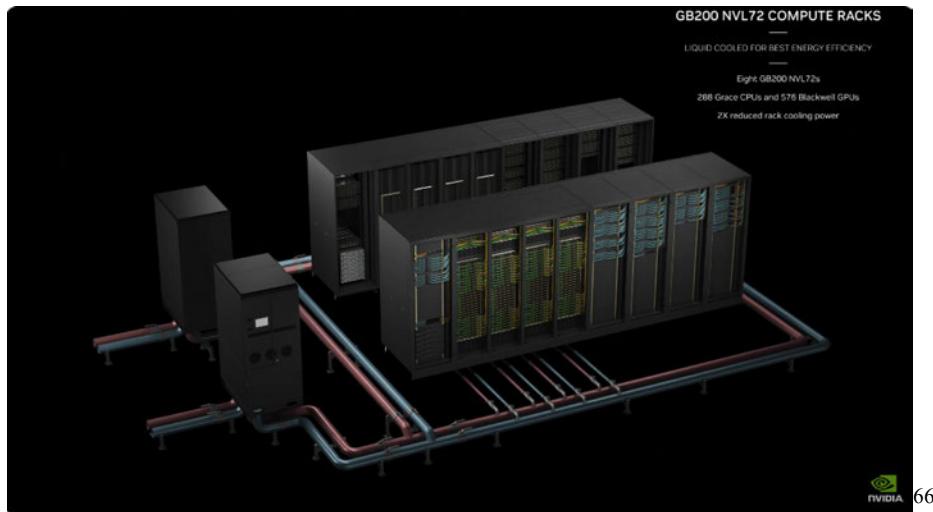
<sup>63</sup> <https://www.nvidia.com/en-us/networking/products/data-processing-unit>.

226. On information and belief, at least by March 2024, NVIDIA advertised that the NVIDIA Blackwell Architecture will be available in Fall 2024, and described how it is “breaking barriers in accelerated computing and generative AI.”<sup>64</sup> Among its “Technological Breakthroughs” are the “NVLink and NVLink Switch” DPUs:

### NVLink and NVLink Switch

Unlocking the full potential of exascale computing and trillion-parameter AI models hinges on *the need for swift, seamless communication among every GPU within a server cluster. The fifth-generation of NVIDIA NVLink interconnect can scale up to 576 GPUs to unleash accelerated performance for trillion- and multi-trillion parameter AI models.*

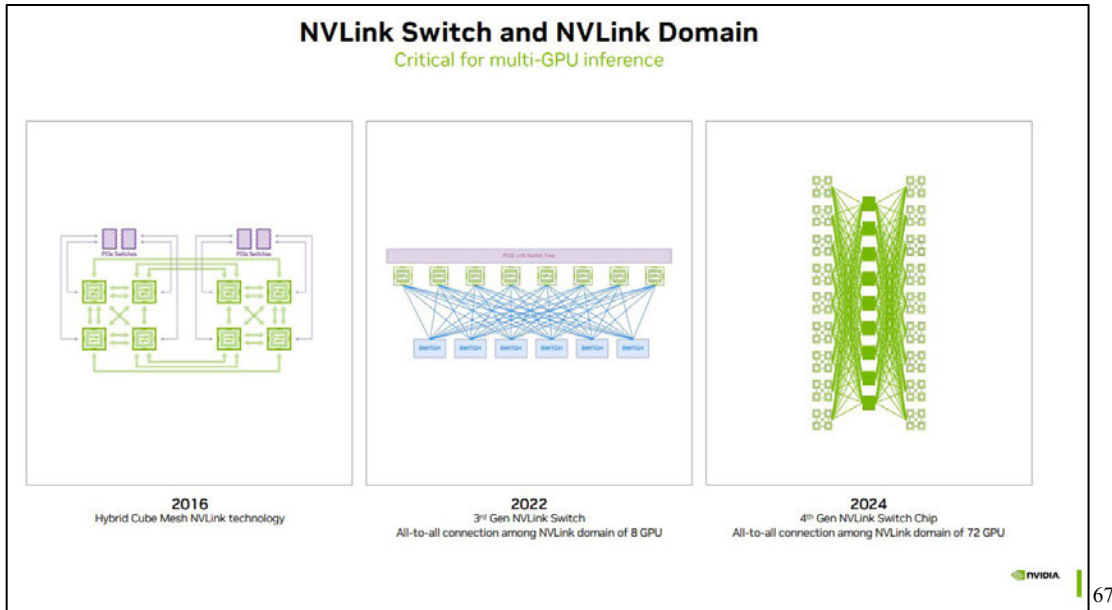
*The NVIDIA NVLink Switch Chip enables 130TB/s of GPU bandwidth in one 72-GPU NVLink domain (NVL72) and delivers 4X bandwidth efficiency with NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)<sup>TM</sup> FP8 support. The NVIDIA NVLink Switch Chip supports clusters beyond a single server at the same impressive 1.8TB/s interconnect. Multi-server clusters with NVLink scale GPU communications in balance with the increased computing, so NVL72 can support 9X the GPU throughput than a single eight-GPU system.*<sup>65</sup>



<sup>64</sup> <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture>.

<sup>65</sup> <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture>.

<sup>66</sup> <https://youtu.be/Y2F8yisiS6E?t=1718> (28:27–29:05).



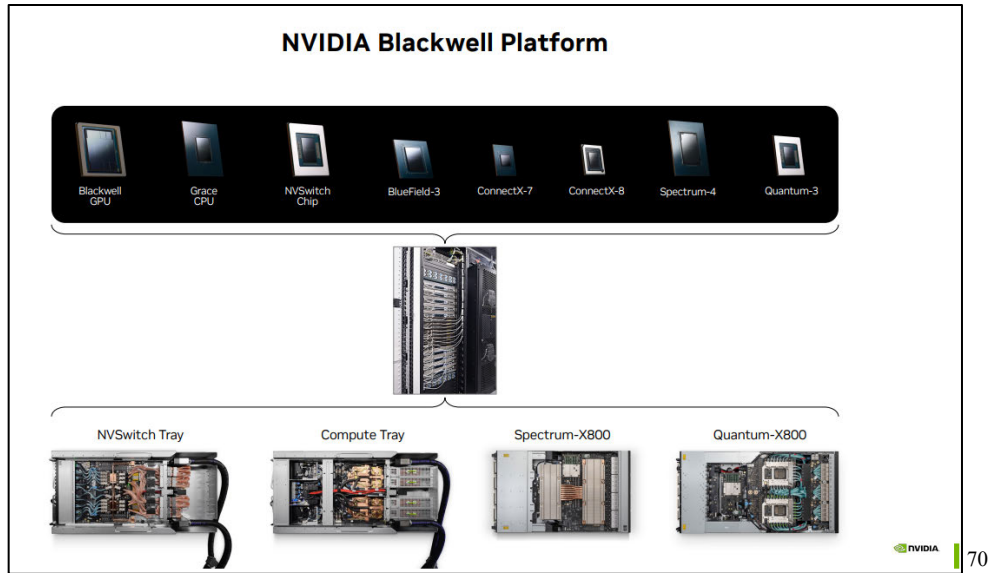
227. NVIDIA described the Blackwell GPU-enabled server computer systems as providing “[s]ignificant performance and power improvements for AI training, inference and accelerated computing.”<sup>68</sup> The Blackwell Platform includes “multiple NVIDIA chips, including the Blackwell GPU, Grace CPU, BlueField data processing unit, ConnectX network interface card, NVLink Switch, Spectrum Ethernet switch and Quantum InfiniBand switch.”<sup>69</sup>

<sup>67</sup> [https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024\\_page\\_23](https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_23).

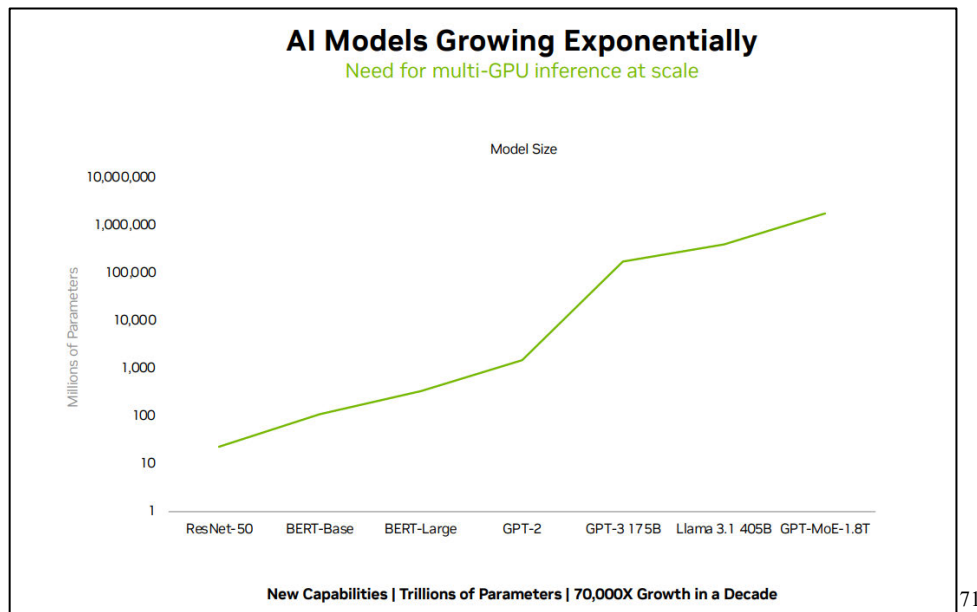
<sup>68</sup> [https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024\\_page\\_32](https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_32).

<sup>69</sup> <https://blogs.nvidia.com/blog/hot-chips-2024>.





228. As Dr. Dalal predicted the explosion of data as a cloud computing challenge, NVIDIA illustrated how the size of large models for ML/AI has grown exponentially to trillions of parameters:



<sup>70</sup> [https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024\\_page\\_06](https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_06).

<sup>71</sup> [https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024\\_page\\_20](https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_20).

229. NVIDIA also described the capabilities of its new Blackwell GPU-enabled server computer system called the “NVIDIA GB200 NVL72” system and its ability to “unlock real-time, trillion-parameter models”:

*The NVIDIA GB200 NVL72 connects 36 GB200 Grace Blackwell Superchips with 36 Grace CPUs and 72 Blackwell GPUs in a rack-scale design. The GB200 NVL72 is a liquid-cooled solution with a 72-GPU NVLink domain that acts as a single massive GPU—delivering 30X faster real-time inference for trillion-parameter large language models.*<sup>72</sup>

230. NVIDIA published a technical brief entitled “NVIDIA Blackwell Architecture Technical Brief: Powering the New Era of Generative AI and Accelerated Computing,” illustrating the capabilities of NVIDIA’s Blackwell Platform<sup>73</sup>:

### Pioneering AI Innovation

In the rapidly evolving landscape of AI and [large language models](#) (LLMs), the pursuit of [real-time performance](#) and [scalability](#) is paramount. From healthcare to automotive industries, organizations are diving deeper into the realms of [generative AI](#) and [accelerated computing](#) solutions. This surge in demand for generative AI solutions is catalyzing a growing need to accommodate ever-expanding model sizes and complexities across enterprises.

Enter [NVIDIA Blackwell GPU architecture](#), the world’s largest GPU, built with the specific purpose of handling data center-scale generative AI workflows with up to 25X the [energy efficiency](#) of the prior NVIDIA Hopper GPU generation.

<sup>72</sup> <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture>.

<sup>73</sup> <https://resources.nvidia.com/en-us-blackwell-architecture/blackwell-architecture-technical-brief>, at 4.

## NVIDIA Blackwell GPU and Superchip Overview

[Large language models](#) (LLMs) require immense computational power for real-time performance. The computational demands of LLMs also translate into higher energy consumption as more and more memory, accelerators, and servers are required to fit, train, and infer from these models. Organizations aiming to deploy LLMs for real-time inference must grapple with these challenges.

The NVIDIA Blackwell architecture and portfolio of products are designed to address the needs of ever-increasing AI model sizes and parameters with a long list of new innovations, including a new second-generation Transformer Engine.

The NVIDIA Blackwell architecture was named to honor [David H. Blackwell](#), an amazing and inspiring American mathematician and statistician known for the Rao-Blackwell Theorem, and many contributions and advancements in probability theory, game theory, statistics, and dynamic programming.

With NVIDIA Blackwell products, every enterprise can use and deploy state-of-the-art LLMs with affordable economics, optimizing their business with the benefits of generative AI. At the same time, NVIDIA Blackwell products enable the next era of generative AI models, supporting multi-trillion parameter models with real-time performance, something unattainable without Blackwell's innovations.

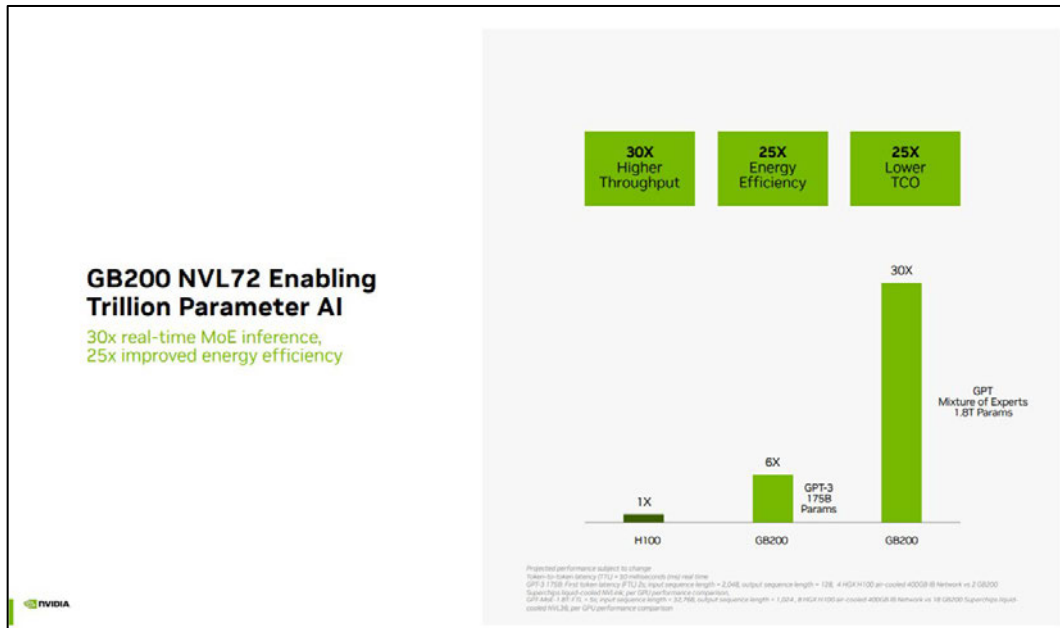
231. On March 18, 2024, NVIDIA issued a press release entitled “NVIDIA Blackwell Platform Arrives to Power a New Era of Computing.” The high points of the press release included the following:

- New Blackwell GPU, NVLink and Resilience Technologies Enable Trillion-Parameter-Scale AI Models
- New Tensor Cores and TensorRT-LLM Compiler Reduce LLM Inference Operating Cost and Energy by up to 25x
- *New Accelerators Enable Breakthroughs in Data Processing*, Engineering Simulation, Electronic Design Automation, Computer-Aided Drug Design and Quantum Computing
- Widespread Adoption by Every Major Cloud Provider, Server Maker and Leading AI Company.<sup>74</sup>

---

<sup>74</sup> <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

232. In the press release, NVIDIA stated: “Powering a new era of computing, NVIDIA today announced that the NVIDIA Blackwell platform has arrived — enabling organizations everywhere to *build and run real-time generative AI on trillion-parameter large language models at up to 25x less cost and energy consumption than its predecessor [the Hopper platform]*.”<sup>75</sup>



233. The press release also highlighted the NVIDIA Blackwell system’s “revolutionary technologies,” including the “Fifth Generation NVLink” with NVLink Switch DPUs:

Fifth-Generation NVLink — To accelerate performance for multitrillion-parameter and mixture-of-experts AI models, the latest iteration of NVIDIA NVLink® delivers groundbreaking 1.8TB/s bidirectional throughput per GPU, *ensuring seamless high-speed*

<sup>75</sup> <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

<sup>76</sup> <https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024> page 26.

*communication among up to 576 GPUs for the most complex LLMs.*<sup>77</sup>

234. The press release further emphasized the importance of the NVLink in the Blackwell system:

**A Massive Superchip**

The NVIDIA GB200 Grace Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s ultra-low-power NVLink chip-to-chip interconnect.

For the highest AI performance, GB200-powered systems can be connected with the NVIDIA Quantum-X800 InfiniBand and Spectrum™-X800 Ethernet platforms, also announced today, which deliver advanced networking at speeds up to 800Gb/s.

*The GB200 is a key component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink.*

Additionally, GB200 NVL72 includes NVIDIA BlueField®-3 data processing units to enable cloud network acceleration, composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds. *The GB200 NVL72 provides up to a 30x performance increase compared to the same number of NVIDIA H100 Tensor Core GPUs for LLM inference workloads and reduces cost and energy consumption by up to 25x.*

*The platform acts as a single GPU* with 1.4 exaflops of AI performance and 30TB of fast memory, and is a building block for the newest DGX SuperPOD.<sup>78</sup>

235. The press release stated “Blackwell-based products will be available from partners starting later this year” and emphasized NVIDIA’s large global network of “Blackwell Partners”:

*AWS, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure* will be among the first cloud service providers to

---

<sup>77</sup> <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

<sup>78</sup> <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.

offer Blackwell-powered instances, as will *NVIDIA Cloud Partner program companies Applied Digital, CoreWeave, Crusoe, IBM Cloud, Lambda and Nebius*. *Sovereign AI clouds* will also provide Blackwell-based cloud services and infrastructure, including *Indosat Ooredoo Hutchinson, Nexgen Cloud, Oracle EU Sovereign Cloud, the Oracle US, UK, and Australian Government Clouds, Scaleway, Singtel, Northern Data Group's Taiga Cloud, Yotta Data Services' Shakti Cloud and YTL Power International*.

GB200 will also be available on NVIDIA DGX™ Cloud, an AI platform co-engineered with leading cloud service providers that gives enterprise developers dedicated access to the infrastructure and software needed to build and deploy advanced generative AI models. AWS, Google Cloud and Oracle Cloud Infrastructure plan to host new NVIDIA Grace Blackwell-based instances later this year.

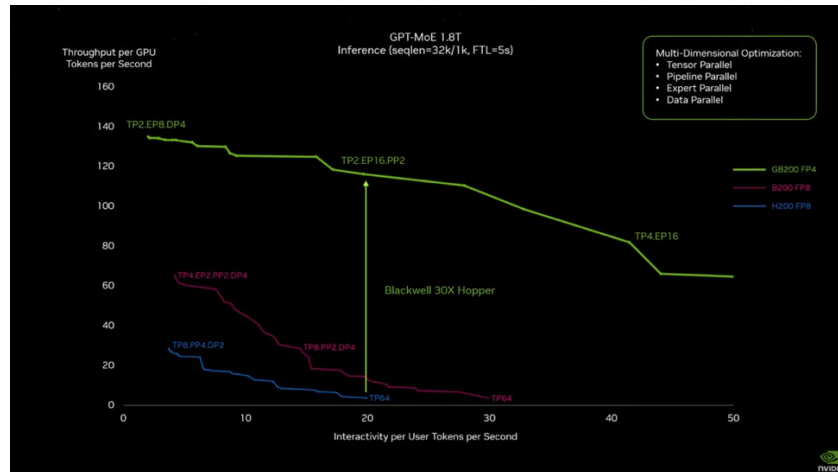
*Cisco, Dell, Hewlett Packard Enterprise, Lenovo and Supermicro* are expected to deliver a wide range of servers based on Blackwell products, as are *Aivres, ASRock Rack, ASUS, Eviden, Foxconn, GIGABYTE, Inventec, Pegatron, QCT, Wistron, Wiwynn and ZT Systems*.

Additionally, a growing network of software makers, including *Ansys, Cadence and Synopsys* — global leaders in engineering simulation — will use Blackwell-based processors to accelerate their software for designing and simulating electrical, mechanical and manufacturing systems and parts. Their customers can use generative AI and accelerated computing to bring products to market faster, at lower cost and with higher energy efficiency.<sup>79</sup>

236. Contemporaneously with the above press release, NVIDIA's CEO Jensen Huang gave a keynote presentation at NVIDIA's GPU Technology Conference (GTC) in March 2024, where he boasted about the Blackwell Platform and the importance of the NVLink Switch DPUs for collective communication and in-network computing in training large language models for ML/AI:

---

<sup>79</sup> <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>.



“Let’s take a look at [the inference capability] of Blackwell compared to Hopper, and this is the extraordinary thing. In one generation, because we created a system that is designed for trillion parameter generative AI, *the inference capability of Blackwell is off the charts. And in fact, it is some 30 times Hopper for large language models like Chat GPT and others like it . . . And very importantly the NVLink Switch, and the reason for that is because all these GPUs have to share the results, partial products [of AI model training operations] . . . Whenever they do all-to-all, all-gather, whenever they communicate with each other, that NVLink Switch is communicating almost 10 times faster than what we could do in the past using the fastest networks.*

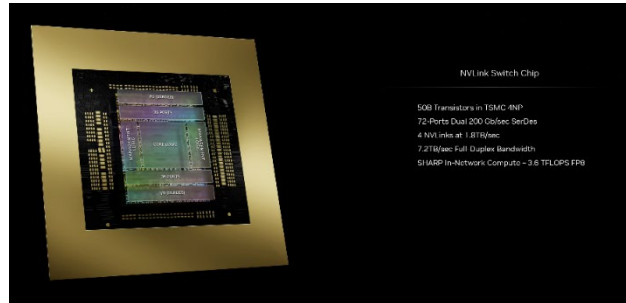
And so Blackwell is going to be just an amazing system for generative AI. And in the future, data centers are going to be thought of—as I mentioned earlier—as an AI Factory. And an AI factory’s goal in life . . . in this industrial revolution is the generation of intelligence, and this ability is super, super important.”<sup>80</sup>

237. During this keynote, NVIDIA further touted the use of the NVLink Switch DPUs for collective communication between GPUs and for in-network computing:

<sup>80</sup> GTC March 2024 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=Y2F8yisiS6E&t=3403s> (56:43–59:06).



“The rate at which we’re advancing computing is insane, and it’s still not fast enough so we built another chip. This [DPU] chip is just an incredible chip. We call it the NVLink Switch. . . . We can have every single GPU talk to every other GPU at full speed at the same time. That’s insane.”<sup>81</sup>



238. With this technology, enabled by at least Xockets’ New Cloud Fabric Patents, NVIDIA emphasized the success NVIDIA’s Blackwell AI factories will have among the world’s AI companies:



***“Blackwell is going to be ramping to the world’s AI companies [linked in the press release above and shown in this video] of which there are so many now doing amazing work in different modalities; every CSP [Cloud Service Provider] is geared up, all the OEMs and ODMs, regional clouds, sovereign AIs, and Telcos all over the world are signing up to launch with Blackwell. Blackwell would be the most successful product launch in our history.”<sup>82</sup>***

<sup>81</sup> GTC March 2024 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=Y2F8yisiS6E&t=2621s> (43:41–44:30).

<sup>82</sup> GTC March 2024 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=Y2F8yisiS6E&t=3609s> (1:00:09–49).



**C. NVIDIA’S KNOWLEDGE OF THE XOCKETS PATENTS**

239. NVIDIA has had actual notice of Xockets’ New Cloud Processor and New Cloud Fabric Patents at least since February 2022.

240. NVIDIA is and was well aware of Xockets’ breakthrough invention of DPU computing architecture and switching fabric as detailed herein, including the invention of a virtual switch for implementing programmable hardware acceleration in the network for cloud offload of data-intensive distributed computing operations independent of host CPUs/GPUs in servers and the invention of a switching fabric for connecting CPUs/GPUs independent of the existing cloud network for cloud offload of data-intensive distributed computing operations, including training large models for AI.

241. On January 27, 2022, Xockets’ co-founder, Dr. Dalal, communicated with NVIDIA’s Brad Genereaux (Global Lead, Healthcare Alliances) and asked for an introduction to “Nvidia legal IP Counsel” in order to discuss “some very strategic IP” that “Nvidia would be interested in acquiring.” Dr. Dalal sought to present NVIDIA the opportunity to acquire exclusive rights to Xockets’ patent portfolio. After making an internal inquiry, Mr. Genereaux ultimately connected Dr. Dalal with Gady Rosenfeld on February 4, 2022.

242. Mr. Rosenfeld was “leading the DPU segment in the NVIDIA field organization” at that time. Indeed, Mr. Rosenfeld’s LinkedIn profile reflects that he has been NVIDIA’s Vice President, DPU Business since July 2021 and remains in that role today. Dr. Dalal and Mr. Rosenfeld had a Teams meeting on February 10 to discuss Xockets and its IP. Dr. Dalal walked Mr. Rosenfeld through exemplary claim charts and explained the nature of Xockets’ patented technology. Mr. Rosenfeld indicated during that meeting that the technology was “extremely interesting.” Later that same day, Dr. Dalal emailed Mr. Rosenfeld sample claim charts and a list of Xockets’ then-current patent list covering breakthrough DPU technologies essential to AI,

which included the New Cloud Processor Patents (the '209, '924, and '350 Patents) and the New Cloud Fabric Patents (the '297, '161, '092, 'and '640 Patents).

243. Mr. Rosenfeld told Dr. Dalal that he would discuss Xockets' patent portfolio with NVIDIA's legal department and then follow up on next steps.

#### **IV. MICROSOFT'S USE OF XOCKETS' TECHNOLOGY**

244. Microsoft infringes the New Cloud Processor Patents at least through its use of NVIDIA BlueField DPUs and ConnectX DPUs (ConnectX-5 and later versions), and infringes the New Cloud Fabric Patents at least through its use of NVIDIA NVLink Switch DPUs, in its server systems in its Microsoft Azure Cloud platform, including NVIDIA's Hopper and Blackwell GPU-enabled server computer systems available in DGX, HGX, MGX, and other configurations (hereinafter, the "Microsoft Accused Products").

245. Microsoft holds the dominant market position in GPU-enabled generative AI platforms via its agreements with leading generative AI model companies, including OpenAI, and its agreements with NVIDIA. Microsoft is in the process of creating and/or maintaining a monopoly in this field.

246. NVIDIA and Microsoft have formed a cartel to monopolize GPU-enabled generative artificial intelligence by controlling the equipment and platforms necessary to access this capability. For example, Microsoft and NVIDIA publicly tout that Microsoft is gaining first access to its GPU-enabled generative AI servers and that Microsoft is embedding NVIDIA technology into the GPU-enabled generative AI platform market it dominates. This creates a self-reinforcing cycle in which users who desire this capability have no choice but to use NVIDIA and Microsoft because of their dominant combined position.

**Microsoft and NVIDIA Announce Major Integrations to Accelerate Generative AI for Enterprises Everywhere**

- Microsoft Azure to Adopt NVIDIA Grace Blackwell Superchip to Accelerate Customer and First-Party AI Offerings
- NVIDIA DGX Cloud’s Native Integration with Microsoft Fabric to Streamline Custom AI Model Development with Customer’s Own Data
- NVIDIA Omniverse Cloud APIs First on Azure Power Ecosystem of Industrial Design and Simulation Tools
- Microsoft Copilot Enhanced with NVIDIA AI and Accelerated Computing Platforms
- New NVIDIA Generative AI Microservices for Enterprise, Developer and Healthcare Applications Coming to Microsoft Azure AI<sup>83</sup>

247. NVIDIA advertises that Microsoft Azure uses NVIDIA’s “DGX Cloud” system.<sup>84</sup>

248. Both NVIDIA and Microsoft publicly promote Microsoft’s use of the accused NVIDIA DPU-enabled systems that copy Xockets’ technology. For example, NVIDIA advertises that “Microsoft Azure and NVIDIA are empowering enterprises to achieve new levels of innovation. With NVIDIA’s *full-stack accelerated computing platform* combined with Microsoft’s global-scale, simplified infrastructure management, enterprises can transform their businesses.”<sup>85</sup>

---

<sup>83</sup> <https://nvidianews.nvidia.com/news/microsoft-nvidia-generative-ai-enterprises>.

<sup>84</sup> <https://www.nvidia.com/en-us/data-center/dgx-cloud>; <https://azuremarketplace.microsoft.com/en-us/marketplace/apps/nvidia.dgx-cloud?tab=Overview> (referring to “NVIDIA DGX™ Cloud on Microsoft Azure”).

<sup>85</sup> <https://www.nvidia.com/en-us/data-center/dgx-cloud>.

249. NVIDIA also advertises that it is “partnering with Microsoft to accelerate the development and deployment of generative AI across Microsoft Azure, Azure AI services, Microsoft Fabric, and Microsoft 365.”<sup>86</sup>

250. NVIDIA also advertises that Microsoft Azure uses NVIDIA’s BlueField DPUs<sup>87</sup>:



#### A. MICROSOFT’S PRAISE OF XOCKETS’ PATENTED TECHNOLOGY

251. Microsoft has described building out its Azure infrastructure as “[t]he most important thing we’ve done over the last four years”:

*“The most important thing is what we’ve done over the last four years [since 2019] is to actually build out the core infrastructure on which OpenAI is built. I mean, these large models, the training infrastructure and the [ML/AI] inference infrastructure doesn’t look like just vanilla cloud, right? So we have had to essentially evolve*

<sup>86</sup> <https://www.nvidia.com/en-us/events/microsoft-build>.

<sup>87</sup> GTC 2023 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=DiGB5uAYKAg&t=1884s> (31:24–31:39).

*Azure [with NVIDIA] to be pretty specialized AI infrastructure. . . .”<sup>88</sup>*

252. In addition, Microsoft has boasted about the benefits the technology brings to Microsoft Azure:

*“Together with NVIDIA, we are making the promise of AI real, helping drive new benefits and productivity gains for people and organizations everywhere,”* said Satya Nadella, chairman and CEO, Microsoft. *“From bringing the GB200 Grace Blackwell processor to Azure, to new integrations between DGX Cloud and Microsoft Fabric, the announcements we are making today will ensure customers have the most comprehensive platforms and tools across every layer of the Copilot stack, from silicon to software, to build their own breakthrough AI capability.”<sup>89</sup>*

253. Microsoft has emphasized the significance of its collaboration with NVIDIA in delivering “state-of-the-art AI capabilities for every enterprise on Microsoft Azure”:

*“AI is fueling the next wave of automation across enterprises and industrial computing, enabling organizations to do more with less as they navigate economic uncertainties,”* said Scott Guthrie, executive vice president of the Cloud + AI Group at Microsoft. *“Our collaboration with NVIDIA unlocks the world’s most scalable supercomputer platform, which delivers state-of-the-art AI capabilities for every enterprise on Microsoft Azure.”<sup>90</sup>*

254. Microsoft has boasted that with NVIDIA, it is providing “the most powerful AI supercomputer” in the world to its customers:

*“The next wave of computing is being born, between next-generation immersive experiences and advanced foundational AI models, we see the emergence of a new computing platform,”* said Satya Nadella, chairman and CEO of Microsoft. *“Together with NVIDIA, we’re focused on both building out services that bridge the digital and physical worlds to automate, simulate and predict*

---

<sup>88</sup> Why Microsoft’s CEO is Ready to Take on Google with ChatGPT, <https://www.youtube.com/watch?v=QinFy0RFD8&t=163s> (2:43–3:05).

<sup>89</sup> <https://news.microsoft.com/2024/03/18/microsoft-and-nvidia-announce-major-integrations-to-accelerate-generative-ai-for-enterprises-everywhere>.

<sup>90</sup> <https://nvidianews.nvidia.com/news/nvidia-microsoft-accelerate-cloud-enterprise-ai>.

*every business process, and bringing the most powerful AI supercomputer to customers globally.”*<sup>91</sup>

**B. MICROSOFT’S KNOWLEDGE OF THE XOCKETS PATENTS**

255. Microsoft is and was well aware of Xockets’ breakthrough invention of DPU computing architecture and switching fabric as detailed herein, including the invention of a virtual switch for implementing programmable hardware acceleration in the network for cloud offload of data-intensive distributed computing operations independent of host CPUs/GPUs in servers and the invention of a switching fabric for connecting CPUs/GPUs independent of the existing cloud network for cloud offload of data-intensive distributed computing operations, including training large models for AI.

256. In 2015 at the most important big data technology conference in the world Xockets presented its technology.

257. Xockets and Microsoft began discussing the potential benefits to Microsoft of Xockets’ technology in May 2016. In March 2017, after a large company expressed an interest in acquiring Xockets, Xockets’ Dan Alvarez reached out to Microsoft’s Ulrich Homann (Corporate Vice President, Cloud and AI) and Jim Brisimitzis (General Manager, Cloud Developer Relations) with a call for bids. The call for bids provided an overview of Xockets’ technology and the fact that Xockets already had a large number of issued patents on the technology:

---

<sup>91</sup> <https://nvidianews.nvidia.com/news/nvidia-and-microsoft-to-bring-the-industrial-metaverse-and-ai-to-hundreds-of-millions-of-enterprise-users-via-azure-cloud>.

## WHAT DOES XOCKETS DO?

### XOCKETS DESIGNS THE XSTREAM APPLIANCE

Public cloud providers, web-scale services companies, and OEMs can directly create new, unique, and powerful **hardware-accelerated services, just by programming software.**

#### How?

The XStream contains the worlds first physical, streaming processors. Our appliance inserts stream processing into the spine of clusters making the most difficult Machine Learning, batch Map-Reduce, or in-memory streaming analytics applications thousands of times faster, using a fraction of resources.

CONFIDENTIAL AND PROPRIETARY

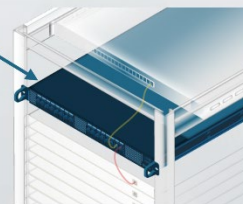


## XSTREAM APPLIANCE

320 Gb/s to 2.2 Tb/s of streaming processing

- >1000x Faster BigData computing
- >1000x Faster BigData repartitioning / sort
- >1000x Faster database joins
- >10x ROI in Machine learning over GPUs

- Less than 2x cost of server
- No change to users' code
- Available for Hadoop and Spark demonstrations today



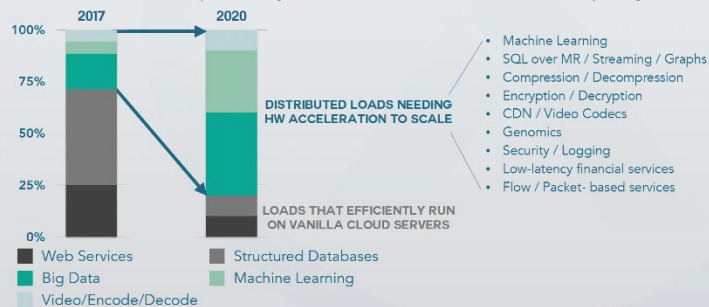
**TOP OF RACK, BUMP-IN-WIRE DEPLOYMENT**  
XStream inserts reconfigurable, streaming processors into the switching spine of clusters

CONFIDENTIAL AND PROPRIETARY



## WHY SPINE PROCESSING?

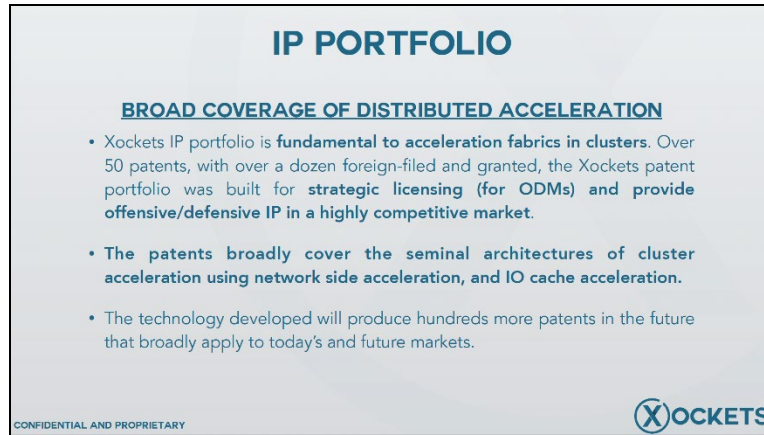
Cloud workloads experiencing a seismic transition in distributed computing.



CONFIDENTIAL AND PROPRIETARY







**Exhibit 8:** “Xockets in a Nutshell” from Microsoft emails, at 2–4, 8.

258. In response, Mr. Homann responded that the “concept resonates and the team would like to understand in more depth.” Mr. Homann directed Xockets to interface with Saurabh Kulkarni (Director of Engineering, Cloud and AI System Technologies) and Kushagra Vaid (VP and Distinguished Engineer, Azure Infrastructure). Ultimately, Dr. Dalal had a discussion with Mr. Kulkarni and Tanj Bennett (Partner SDE) on March 22, 2017, so they could “get a technical overview of key Xockets technologies in the hardware acceleration space.” Thereafter, Mr. Kulkarni informed Dr. Dalal that he was reaching out to folks from Microsoft’s “big data and machine learning teams” in order to make an introduction. As discussed further below, instead of further engaging with Xockets regarding its technology, Microsoft just took Xockets’ technology without paying for it. And when Xockets subsequently approached Microsoft about taking a license, Microsoft formed a buyers’ cartel with NVIDIA whereby both of them agreed not to compete for the license or purchase of Xockets’ technology, but instead to negotiate only through RPX in order to obtain a price below what would have been obtained under normal, non-collusive market conditions.



## V. REPRESENTATIVE BENEFITS OF XOCKETS' PATENTED TECHNOLOGY

259. Implementing Xockets' patented inventions through DPUs for cloud offload processing provides multiple benefits, including: Total Cost of Ownership ("TCO") Savings and Accelerated Performance.

260. **Increased TCO Savings:** With respect to TCO Savings—as described above, data centers are under increasing pressure to keep operating costs down. Innovations which provide the opportunity to purchase less hardware equipment (such as CPUs or GPUs), and use less power (thus saving on power consumption costs), are hugely valuable to data centers. Xockets' inventions provide these exact benefits.

261. DPUs for cloud offload processing enhance server efficiency by offloading data intensive infrastructure tasks involved in managing the flows of packets in a cloud data center, thereby freeing up valuable CPU or GPU cycles. Without Xockets' patented inventions, these infrastructure tasks were previously performed by the server processors, such as CPUs or GPUs. By offloading these tasks to DPUs, freeing up CPU or GPU cycles, cloud data centers can do more processing with less hardware, maximizing their return on investment. As one example, NVIDIA estimates that a single DPU can replace 300 CPUs.<sup>92</sup>

262. Cloud operators can exchange these TCO savings for denser, higher performance data centers that can produce higher revenues and profits by running more customer applications and services and accelerating their performance.

263. NVIDIA has performed studies corroborating these benefits. For example, in a November 2022 White Paper titled "DPU Power Efficiency," attached hereto as **Exhibit 9**,

---

<sup>92</sup> <https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3>.

NVIDIA estimated that DPUs for cloud offload processing can “reduce server power consumption up to 30%. . . . plus additional savings in cooling, power deliver, rack space, and server capital costs.” **Exhibit 9** at 4. NVIDIA also stressed the importance of power savings, explaining “[n]ow that most data centers can be brought online rapidly and offer high levels of availability and compute density, improving power consumption and reducing associated power costs have become top goals both for optimizing existing data centers and designing new ones.” *Id.*

264. NVIDIA performed several tests which show that offloading different categories of tasks result in significant TCO Savings. As one specific example, NVIDIA found that the 3-year TCO savings from offloading **only** IPsec encryption/decryption to a BlueField DPU (such as used in Cloud VPN network overlay services) was approximately **\$3,207** per server in 3-year TCO savings for **cloud security offload** (based on a savings of \$26.3 million across 8,200 servers, as shown in the table below). *Id.* at 21, Table 7. NVIDIA explained: “We see significant two-way savings from the offload and acceleration capabilities of the BlueField DPU. The offload frees up CPU cores allowing fewer servers to be deployed, reducing CapEx. The lower number of servers and lower per-server power consumption combine to reduce OpEx substantially. The result is a substantial savings of \$26M over three years in a large data center with 10,000 servers.” *Id.*

Table 7. TCO calculation from offloading IPsec encryption/decryption to a BlueField DPU, for a large data center with 10,000 servers.		
Large Data Center TCO	Servers without DPU	Servers with DPU Offload
Servers needed	10,000	8,200 (18% reduction)
Cost per server	\$10,500 (no DPU)	\$12,000 (with DPU) <sup>10</sup>
Total Server CapEx	\$105,000,000	\$98,400,000 (\$6.6M / 6.3% savings)
Power use per server	728W (0.728 kW)	481W (247W/34% reduction)
Total power use, 3 years	191,318,400 kWh	103,653,576 kWh (45.8% reduction)
Server power cost (\$0.15/kWh)	\$28,697,760	\$15,548,036 (\$13.1M savings)
Total power cost (PUE=1.5)	\$43,046,640	\$23,322,054 (\$19.7M OpEx savings)
3-year TCO (CapEx + OpEx)	\$148,046,640	\$121,722,054 ( <b>\$26.3M / 17.8% savings</b> )

265. Furthermore, one of NVIDIA's technical blogs estimates the minimum TCO Savings of approximately **\$2,137** per server in minimum TCO savings for *ML/AI collective communication offload* (based on a savings of \$18 million across an estimated 8,422 servers, using NVIDIA's methodology in Table 7 above):

These key advancements enable BlueField-3 to run workloads up to 8x faster while reducing the TCO and delivering data center energy efficiency. For example, Bluefield-3 *offloads HPC/AI MPI collective operations from the CPU, delivering nearly a 20% increase in speed, which translates to \$18 million dollars in cost savings* for large-scale supercomputers.<sup>93</sup>

266. Significantly, each of these TCO savings were calculated based on the offload of just security and collective communication offload. In reality, as described above, DPUs allow for the offload of other cloud virtualization, network, storage, and security tasks. NVIDIA CEO Jensen Huang explained that these tasks in total “*can consume nearly half of the data center's CPU cores and associated power.*”<sup>94</sup>

267. Huang confirmed that offloading all data-intensive cloud infrastructure services to DPUs can reduce approximately 50% of a cloud data center's power usage (OpEx) and server requirements (CapEx). By using NVIDIA's own method of calculating TCO Savings, the calculated 3-year TCO Savings of **\$15,457 per server in TCO savings** (based on a savings of approximately \$77.2 million across an estimated 5,000 servers, using NVIDIA's methodology in Table 7 above):

---

<sup>93</sup> <https://developer.nvidia.com/blog/power-the-next-wave-of-applications-with-Nvidia-BlueField-3-dpus>.

<sup>94</sup> GTC 2023 Keynote with NVIDIA CEO Jensen Huang, <https://www.youtube.com/watch?v=DiGB5uAYKAg&t=1836s> (30:36–49).

<b>Large data center TCO</b>	<b>Servers without DPU</b>	<b>Servers with DPU offload</b>
Servers needed	10,000	5,000
Cost per server	\$10,500	\$12,000
Total server CapEx	\$105,000,000	\$60,000,000
Power use per server	728 Watts (W)	364 Watts (W)
Total power use, 3 years	191,318,400 kWh	47,829,600 kWh
Server power cost (\$0.15/kWh)	\$28,697,760	\$7,174,440
Total power cost (PUE = 1.5)	\$43,046,640	\$10,761,660
3-year TCO (CapEx + OpEx)	\$148,046,640	\$70,761,660
<b>3-year TCO Savings (\$/server)</b>	-	<b>\$77,284,980 (\$15,457/server)</b>
3-year TCO Savings (%)	-	52.2%

268. Xockets anticipates that the actual cost savings will be much greater.

269. NVIDIA has explained that by using DPUs for offloading cloud processing “[t]here will typically also be additional savings from the ability to run more revenue-generating workloads as well on each server thanks to the CPU cycles freed up by the networking offload. Deploying DPU offloads in servers usually allows each server to perform more work (more connections, more virtual machines, more users, etc.). This results in a large CapEx savings because fewer servers are needed, as well as a significant OpEx savings because fewer servers consume less power, floor space, and other data center resources (cooling, power distribution, management).” **Exhibit 9** at 20.

270. **Accelerated Performance:** The use of DPUs for cloud offload processing also accelerates the performance of cloud applications by enabling additional compute cycles, which results in reducing latency and an enhanced end user experience. The improved application responsiveness and reliability have a direct impact on customer satisfaction, user engagement, and higher transaction volumes and prices, all of which contribute to increased prices and revenue.

271. For example, on the low end, NVIDIA has estimated that “virtualization, networking, storage, security, management, and provisioning” can “consume up to 30% of the

processor cycles.” *Id.* at 8. By using DPUs to free up those cycles, performance is improved and server processor cores are able to run the types of applications they do best. *Id.*

272. Each offload use therefore results in accelerated application performance and reduced latency which further leads to enhanced customer satisfaction, increased user engagement, and higher market share/prices, which in turn leads to increased revenues. On information and belief, NVIDIA accordingly charges higher prices for products that allow for increased acceleration of the performance of cloud applications. On information and belief, this accelerated performance is estimated to provide public cloud providers like Microsoft at least ***\$30,000 per server in increased revenue*** over the 3-year lifespan of a server in cloud data centers.

## VI. RPX’S BUSINESS

273. RPX was founded in 2008 and has more than 450 members, including NVIDIA and Microsoft. RPX’s website explains:

RPX Corporation brings companies together from throughout the world to solve patent risks that they face in common. Our conviction is that solving such problems once for many companies can achieve a faster, better, and less expensive resolution than might otherwise be achieved by each company acting alone. To this end, we offer a platform that includes defensive buying of patent rights, acquisition syndication, patent intelligence, insurance services, and advisory services.

Our pioneering approach combines principal capital, deep patent expertise, and client contributions to generate enhanced patent buying power. By efficiently acquiring rights to problematic patents, we help to mitigate and manage the risk of potential patent assertions for our growing client network.

95

274. RPX previously touted on its website (language that has since been removed) that “[i]n effect, RPX can buy ‘wholesale’ on behalf of our client network, while our clients otherwise would pay ‘retail’ if transacting on their own.” **Exhibit 10.** The RPX website also previously advertised that “RPX is often able to achieve ‘wholesale’ pricing terms, where we can acquire

---

<sup>95</sup> <https://www.rpxcorp.com/about>.

rights for our members at significantly reduced cost relative to what the NPE might charge an individual company on its own. RPX believes we have saved our members tens of millions of dollars through these wholesale-priced transactions.” **Exhibit 11**. Despite RPX having removed the language in an effort to hide its anticompetitive behavior, RPX’s business practices remain the same today.

275. RPX’s most recent 10-K filing with the SEC in 2018 explains its mission of interjecting itself as the “essential intermediary” between patent owners and RPX’s members:

Our mission is to reduce risk and cost for corporate legal departments through data-driven decision-making, technology, and market-based solutions. A significant part of that mission is to transform the patent market by establishing RPX as the essential intermediary between patent owners and operating companies and by providing complementary technology-focused discovery services. Our strategy includes the following:

**Exhibit 12**, at 6. RPX’s business practices remain the same today.

276. RPX’s co-founder and former CEO, John Amster, has publicized RPX’s mission, stating that “[w]e think there can be a clearinghouse in this market that can be really quite big and efficient. If every company just decided, ‘We’re going to have a line item in our budget for patents and patent risks, and that line item is going to be the RPX rate card’—i.e. RPX’s subscription rate[.]”

277. Indeed, RPX’s website (which refers to the “RPX Network” as the “world’s leading defensive patent acquisition network”) touts how RPX’s application of “capital to acquire patents rights” leads to “far less cost” for its members, that “[t]here’s safety in numbers” and “huge cost savings, too.”<sup>96</sup>

---

<sup>96</sup> [www.rpxcorp.com/solutions/rpx-network](http://www.rpxcorp.com/solutions/rpx-network).

278. RPX's website also explains how it collaborates with its members and non-members to create anti-competitive buyers' cartels, what it euphemistically calls "syndicated licensing transactions":

In addition to our core patent acquisition service, RPX also facilitates large-scale syndicated licensing transactions that can include non-members and members (who make contributions beyond their regular subscription fees).

97

279. RPX previously highlighted the benefits of these syndicated transactions for its clients on its website, in language that has since been removed, stating that "[o]ur clients see distinct advantages of syndicated purchasing through RPX, as we are uniquely situated to structure transactions that are ultimately less costly and deliver more value to participating clients than if any attempted individual licensing or unilateral purchasing of the portfolios." **Exhibit 10.**

280. RPX has openly acknowledged in its SEC filings that its practices may be illegal and violate competition and antitrust laws, admitting that "[i]t is possible that courts or other governmental authorities will interpret existing laws regulating [] competition and antitrust practices [] in a manner that is inconsistent with our business practices."<sup>98</sup>

## **VII. NVIDIA AND MICROSOFT RESPOND TO XOCKETS' 2024 FUNDRAISING EFFORTS BY CREATING A BUYERS' CARTEL**

281. In early 2024, Xockets engaged in a process to sell or license its technology [REDACTED] [REDACTED] As part of the effort, NVIDIA and Microsoft were approached about the Xockets technology. Specific to NVIDIA, on March 27, 2024, Xockets' representative emailed NVIDIA's Vishal Bhagwati (Head

<sup>97</sup> <http://ir.rpxcorp.com>.

<sup>98</sup> RPX Corporation, S.E.C. Registration Statement (Form S-1), at 17 (Jan. 21, 2011), available at <https://www.sec.gov/Archives/edgar/data/1509432/000119312511012087/ds1.htm>.



of Corporate Development), Timothy Teter (Executive Vice President and General Counsel), David Shannon (EVP, Chief Administrative Officer and Secretary), and Rich Domingo (Director of Intellectual Property) Xockets' information and a proposed NDA. Xockets' representative separately followed up with Mr. Domingo on April 2 and 9 and June 5, and Gady Rosenfeld (NVIDIA's Vice President, DPU Business) on May 2. On April 30, Xockets' representative also sent a follow up email to his original March 27 email to Messrs. Bhagwati, Teter, Shannon, and Domingo. Around the time of these emails or shortly thereafter, NVIDIA interacted with RPX to form a conspiracy with Microsoft to create a buyers' cartel by refusing to negotiate individually with Xockets and instead only negotiating through RPX.

282. Specific to Microsoft, on March 27, 2024, Xockets' representative emailed Microsoft's Christopher Young (Executive Vice President Business Development), Michael Wetter (Corporate Vice President, Corporate Development), and Nicholas Kim (Senior Corporate Counsel, IP Litigation) the teaser and a proposed NDA. He forwarded that email to Microsoft's Steve Bathiche a day later and Brad Smith (President) on April 2. On April 30, Xockets' representative also followed up the original email with Messrs. Young, Wetter, and Kim, and separately with Mr. Smith. Around the time of these emails or shortly thereafter, Microsoft interacted with RPX to form a conspiracy with NVIDIA to create a buyers' cartel by refusing to negotiate individually with Xockets and instead only negotiating through RPX.

283. In May 2024, RPX's CEO, Dan McCurdy, contacted Xockets' representative to have a call and set up a subsequent dinner meeting. During the conversations, Mr. McCurdy made statements to the effect that Mr. McCurdy was being directed by members who were aware of an available portfolio of intellectual property. It was public that Xockets' representative was affiliated with Xockets, and the Xockets portfolio was the only available portfolio that Xockets'

representative was involved with at the time. Mr. McCurdy indicated he would go back to his members to consider next steps.

284. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

285. Given their AI-driven roles and the necessity of Xockets' technology to those roles as set forth above, NVIDIA and Microsoft constitute a large part of the demand for Xockets' patented technology, and this market strength is exacerbated by their combination and agreement with and use of RPX, which counts among its members other companies that further make up the vast majority of the demand for Xockets' patented technology.

286. Since the time that RPX has become involved, Xockets has been prevented from obtaining a fair market price for its patents, which undeniably read on Microsoft's and NVIDIA's products, and neither Microsoft, nor NVIDIA, nor any of RPX's other members will negotiate at all with Xockets, thereby reducing output in the market for Xockets' patents to effectively zero.

#### **VIII. ILLEGAL AGREEMENT BETWEEN THE DEFENDANTS**

287. NVIDIA and Microsoft have agreed with each other and with RPX not to separately negotiate with Xockets and instead to only negotiate jointly via RPX. The individuals who entered into this agreement include but are not limited to those referenced in the preceding paragraphs. This buyers' cartel has allowed for price fixing by pushing the price and output below what would have been agreed to under normal market conditions. Indeed, the very purpose of RPX and the reason for joining RPX is to form groups of potential purchasers who can use collective purchasing

power for technology inputs, something that NVIDIA and Microsoft understood and sought to employ via their unlawful agreement.

288. RPX's public statements, including but not limited to the statements referenced above, evidence that its platform is being employed for the purposes of a buyers' cartel. NVIDIA's and Microsoft's agreement to only negotiate as a buyers' cartel in the market for Xockets' patents has resulted in NVIDIA and Microsoft behaving in a manner contrary to their self-interest as horizontal competitors in that market. By negotiating individually, each would have had the opportunity to obtain a first mover advantage that in a functioning competitive market results in lower pricing and a competitive advantage against each other in the competition for Xockets' patents (including potentially even an exclusive license). This behavior also reflects an agreement to either exercise or accumulate monopsony power and drive license fees and/or purchase costs substantially below market rates and/or to collectively refuse to license and/or purchase at all, which would drive Xockets out of business.

289. Xockets claims two separate bases for violation of the antitrust laws: (1) a conspiracy to restrain trade in violation of § 1 of the Sherman Act by all defendants, and (2) a conspiracy to monopolize (monopsonize) by all defendants in violation of § 2 of the Sherman Act. These antitrust claims are brought under §§ 4 and 16 of the Clayton Act (15 U.S.C. §§ 15 and 26): (a) to recover damages, including treble damages, sustained by Xockets as a result of its being injured in its business and property by reason of defendants' violations of the antitrust laws, particularly Sections 1 and 2 of the Sherman Act (15 U.S.C. §§ 1 and 2), (b) to obtain injunctive relief against threatened loss or damage as a result of such violations, and (c) to recover the expense of bringing and maintaining this action, including reasonable attorneys' fees.

**COUNT I: VIOLATION OF SECTION 1 OF THE SHERMAN ACT BASED ON  
DEFENDANTS' CONSPIRACY IN RESTRAINT OF TRADE**

290. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

291. The facts set forth herein establish that a contract, combination, or conspiracy exists between and among at least RPX, NVIDIA, and Microsoft which restrained and continues to restrain trade or commerce among the several States in the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents.

292. The agreement that NVIDIA and Microsoft act only through RPX to purchase, acquire, or license the Xockets patent portfolio forms the basis for the illegal contract, combination, or conspiracy. This agreement is to restrain output and fix prices below the market competitive price for Xockets' patented technology and/or to drive Xockets out of business.

293. An agreement exists between NVIDIA and Microsoft based on the alleged facts, including that NVIDIA and Microsoft directed and were therefore aware that RPX was negotiating on their behalf with Xockets and NVIDIA and Microsoft's respective refusal to discuss acquisition of the Xockets patent portfolio separately. The behavior of RPX evidencing that it is representing a buyers' cartel made up of at least NVIDIA and Microsoft also supports a reasonable inference that NVIDIA and Microsoft were acting in concert. RPX's public statements, discussed above, also support the existence of the buyers' cartel as they describe an invitation to concerted action, and to have RPX coordinate that concerted action.

294. The conspiracy, by virtue of Defendants' market power, is unreasonably restrictive of competition and Xockets suffered antitrust injury as a result.

295. The relevant product market is the market for purchase, acquisition, or licensing of technology covered by Xockets' patents.

296. Xockets' antitrust injury includes but is not limited to the fact that Defendants' conduct has destroyed the normal market forces that should have made it possible for Xockets to license or sell its technology. As a result of Defendants' conduct, Xockets has been unable to do so at normal market price, which has led to lost revenues and opportunities, and, if this conduct continues, Xockets will be driven out of business.

297. Moreover, Defendants' conduct not only harms competition with respect to the market for Xockets' patents, it also harms competition in the downstream markets for the market for GPU-enabled AI servers, which is controlled by NVIDIA, and the market for GPU-enabled AI platforms, which is controlled by Microsoft. As noted above, NVIDIA controls over 90% the market for GPU-enabled AI servers and Microsoft, through its partnership with Open AI, controls 70% of the market for GPU-enabled AI platforms. By driving down the costs of Xockets' patents, Microsoft and NVIDIA can continue their dominance of these markets. If successful, this will harm invocation and allow NVIDIA and Microsoft to unilaterally increase prices within these markets.

298. Unless restrained by this Court, Defendants' unlawful conspiracy will continue to impose continuous injury and loss on Xockets' ability to sell or license its technology in a market free from such unlawful behavior.

**COUNT II: VIOLATION OF SECTION 2 OF THE SHERMAN ACT BASED ON DEFENDANTS' CONSPIRACY TO CREATE OR MAINTAIN A MONOPSONY**

299. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

300. Defendants have combined or conspired in an attempt to obtain and/or in fact, have obtained monopsony power in the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents, that power was or is being willfully acquired through Defendants'

overt acts done with a specific intent to achieve monopsony power, and had an effect on a substantial amount of interstate commerce.

301. As detailed herein, the Defendants' anticompetitive conduct has eliminated competition between them for licenses to Xockets' patent portfolio with the effect of price fixing and more favorable non-monetary terms than possible in a market unaffected by such anticompetitive conduct.

302. Defendants either accumulated and are maintaining or are accumulating monopsony power over at least the relevant market and have used and are using that power to prevent Xockets from being able to sell or license its patent portfolio at normal market prices and to restrain quantities and fix prices at below the normal market rates.

303. Defendants' monopsony power is being or was willfully and intentionally acquired. The conspiracy amongst Defendants alleged herein was undertaken for the specific purpose of obtaining monopsony power over the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents.

304. Defendants' monopsony power is not the result of good business skill or acumen, instead it is the product of their illegal conspiracy.

305. Unless restrained by this Court, Defendants' unlawful monopsonization of the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents will continue to impose continuous injury and loss on Xockets' ability to sell or license its technology in a market free from such unlawful behavior. The impact being to illegally effect a substantial amount of interstate commerce.

306. Moreover, Defendants' conduct not only harms competition with respect to the market for Xockets' patents, it also harms competition in the downstream markets for the market

for GPU-enabled AI servers, which is controlled by NVIDIA, and the market for GPU-enabled AI platforms, which is controlled by Microsoft. As noted above, NVIDIA controls over 90% the market for GPU-enabled AI servers and Microsoft, through its partnership with Open AI, controls 70% of the market for GPU-enabled AI platforms. By driving down the costs of Xockets' patents, Microsoft and NVIDIA can continue their dominance of these markets. If successful, this will harm invocation and allow NVIDIA and Microsoft to unilaterally increase prices within these markets.

### **COUNT III: INFRINGEMENT OF THE '209 PATENT**

307. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

#### **I. DIRECT INFRINGEMENT**

308. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '209 Patent's claims, including at least Claim 18 and Claim 20, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA BlueField DPUs and ConnectX DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

309. NVIDIA and Microsoft are infringing claims of the '209 Patent, including at least Claim 18 and Claim 20, literally and/or pursuant to the doctrine of equivalents.

310. Claim 18 of the '209 Patent is directed to a server system, comprising:

a plurality of servers interconnected by a network, each server including

a server processor configured to execute an operating system for the server,



at least one computation module, separate from the server processor and coupled to the server processor by at least one bus, the at least one computation module including

first processing circuits mounted on the computation module and configured to

execute header detection on packets received by the server,

classifying received packets by a session identifier, and

operate as a virtual switch to provide packets to circuits on the at least one computation module, and

at least decryption circuits implemented on programmable logic devices and configured to decrypt received packets; wherein

the computation modules execute header detection, classifying of packets, virtual switching of packets, and decryption of packets independent of the server processor of their respective server.

311. Claim 20 of the '209 Patent is directed to:

The server system of claim 18, wherein the at least decryption circuits decrypt the received packets according to a virtual private network (vpn) encryption/decryption protocol.

**A. NVIDIA'S DIRECT INFRINGEMENT**

312. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '209 Patent, including at least Claim 18 and Claim 20.

313. The NVIDIA Accused Products are server systems.

314. NVIDIA's server systems comprise a plurality of servers (e.g., NVIDIA GPU-centric servers) that are interconnected by a network (e.g., NVIDIA's Quantum InfiniBand network and/or a network including a top-of-rack ("TOR") switch).

315. Each of NVIDIA's servers in the NVIDIA Accused Products includes a server processor (e.g., a host processor) configured to execute a host operating system for the server.

316. In addition, each server in the NVIDIA Accused Products includes at least one computation module (e.g., an NVIDIA BlueField DPU) that is separate from the host processor and coupled to the host processor by at least one bus (e.g., a Peripheral Component Interconnect Express (“PCIe”) bus).

317. Further, the computation module of the NVIDIA Accused Products includes first processing circuits (e.g., a ConnectX DPU subsystem) mounted on it.

318. The first processing circuits of the NVIDIA Accused Products are configured to execute header detection on packets received by the server, classify received packets by a session identifier, and operate as a virtual switch to provide packets to circuits on the at least one computation module (e.g., through the pipeline-based programmable eSwitch of the ConnectX DPU).

319. The computation module of the NVIDIA Accused Products includes decryption circuits (e.g., Inline Hardware IPsec/TLS/CT encryption and decryption circuits of the DPUs) implemented on programmable logic devices (e.g., programmable pipelines of hardware accelerators) and configured to decrypt received packets.

320. Lastly, the computation modules of the NVIDIA Accused Products execute header detection, classifying of packets, virtual switching of packets, and decryption of packets independent of the server processor of their respective server (e.g., by the NVIDIA BlueField DPU offloading, accelerating, and isolating the virtual switch implemented on the ConnectX DPU from the CPU of its server).

321. Lastly, the decryption circuits on the computation modules of the NVIDIA Accused Products can decrypt the received packets according to a virtual private network (vpn) encryption/decryption protocol (e.g., for providing virtual private network (VPN) services).

**B. MICROSOFT'S DIRECT INFRINGEMENT**

322. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '209 Patent, including at least Claim 18 and Claim 20.

323. The Microsoft Accused Products are server systems.

324. Microsoft's server systems comprise a plurality of servers (e.g., NVIDIA or other servers) that are interconnected by a network (e.g., NVIDIA's Quantum InfiniBand network and/or a network including a top-of-rack ("TOR") switch).

325. Each of Microsoft's servers in the Microsoft Accused Products includes a server processor (e.g., a host processor) configured to execute a host operating system for the server.

326. In addition, each server in the Microsoft Accused Products includes at least one computation module (e.g., an NVIDIA BlueField DPU) that is separate from the host processor and coupled to the host processor by at least one bus (e.g., a Peripheral Component Interconnect Express ("PCIe") bus).

327. Further, the computation module of the Microsoft Accused Products includes first processing circuits (e.g., a ConnectX DPU subsystem) mounted on it.

328. The first processing circuits of the Microsoft Accused Products are configured to execute header detection on packets received by the server, classify received packets by a session identifier, and operate as a virtual switch to provide packets to circuits on the at least one computation module (e.g., through the pipeline-based programmable eSwitch of the ConnectX DPU).

329. In addition, the computation module of the Microsoft Accused Products includes decryption circuits (e.g., Inline Hardware IPsec/TLS/CT encryption and decryption circuits of the

DPU) implemented on programmable logic devices (e.g., programmable pipelines of hardware accelerators) and configured to decrypt received packets.

330. The computation modules of the Microsoft Accused Products execute header detection, classifying of packets, virtual switching of packets, and decryption of packets independent of the server processor of their respective server (e.g., by the NVIDIA BlueField DPU offloading, accelerating, and isolating the virtual switch implemented on the ConnectX DPU from the CPU of its server).

331. Lastly, the decryption circuits on the computation modules of the Microsoft Accused Products can decrypt the received packets according to a virtual private network (vpn) encryption/decryption protocol (e.g., for providing virtual private network (VPN) services).

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

332. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '209 Patent's claims, including at least Claim 18 and Claim 20, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '209 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '209 Patent, including at least Claim 18 and Claim 20.

333. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '209 Patent, including at least Claim 18 and Claim 20, to induce third

parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that would infringe one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

334. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs, in manners that would infringe one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

335. NVIDIA has had knowledge of the '209 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA BlueField DPU and ConnectX DPU in server systems would infringe Xockets' Asserted Patents, including the '209 Patent. To the extent that NVIDIA lacked actual knowledge of the '209 Patent or its customers' and/or end-users' actual infringement of the '209 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

336. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20, with knowledge of the '209

Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '209 Patent.

337. At a minimum, NVIDIA has had actual notice of the '209 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 18 and Claim 20 of the '209 Patent by its customers and end-users, including Microsoft.

338. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '209 Patent's claims, including at least Claim 18 and Claim 20, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '209 Patent, including at least Claim 18 and Claim 20.

339. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '209 Patent, including at least Claim 18 and Claim 20, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

340. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

**B. MICROSOFT'S INDIRECT INFRINGEMENT**

341. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '209 Patent's claims, including at least Claim 18 and Claim 20, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '209 Patent, including at least Claim 18 and Claim 20.

342. For example, Microsoft sells over 200 Azure products<sup>99</sup> and over 40 Azure cloud solutions<sup>100</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

343. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '209 Patent, including at least Claim 18 and Claim 20, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

344. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products

---

<sup>99</sup> <https://azure.microsoft.com/en-us/products>.

<sup>100</sup> <https://azure.microsoft.com/en-us/solutions>.



(including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

345. Microsoft has had knowledge of the '209 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '209 Patent. To the extent that Microsoft lacked actual knowledge of the '209 Patent or its customers' and/or end-users' actual infringement of the '209 Patent, Microsoft took deliberate actions to avoid learning of those facts.

346. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20, with knowledge of the '209 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '209 Patent.

347. At a minimum, Microsoft has had actual notice of the '209 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 18 and Claim 20 of the '209 Patent by its customers and end-users.

348. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '209 Patent's claims, including at least Claim 18 and Claim 20, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in

this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '209 Patent, including at least Claim 18 and Claim 20.

349. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '209 Patent, including at least Claim 18 and Claim 20, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

350. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '209 Patent, including at least Claim 18 and Claim 20.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

351. NVIDIA has had knowledge of the '209 Patent no later than February 10, 2022.

352. Despite knowing of the '209 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '209 Patent.

353. Despite knowing of the '209 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '209 Patent.

354. At a minimum, NVIDIA has had actual notice of the '209 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '209 Patent, including at least Claim 18 and Claim 20.

355. At a minimum, NVIDIA has also willfully blinded itself to the '209 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '209 Patent but took deliberate steps to avoid learning of its infringement.

356. Therefore, upon information and belief, NVIDIA's infringement of the '209 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

357. Microsoft has had knowledge of the '209 Patent no later than March 22, 2017.

358. Despite knowing of the '209 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '209 Patent.

359. Despite knowing of the '209 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '209 Patent.

360. At a minimum, Microsoft has had actual notice of the '209 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '209 Patent, including at least Claim 18 and Claim 20.

361. At a minimum, Microsoft has also willfully blinded itself to the '209 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft

Accused Products infringed the '209 Patent but took deliberate steps to avoid learning of its infringement.

362. Therefore, upon information and belief, Microsoft's infringement of the '209 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

#### **COUNT IV: INFRINGEMENT OF THE '924 PATENT**

363. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

##### **I. DIRECT INFRINGEMENT**

364. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '924 Patent's claims, including at least Claim 9, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA BlueField DPUs and ConnectX DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

365. NVIDIA and Microsoft are infringing claims of the '924 Patent, including at least Claim 9, literally and/or pursuant to the doctrine of equivalents.

366. Claim 9 of the '924 Patent is directed to a method for providing network overlay services, comprising the steps of:

receiving network packet data from a data source in an offload processor module that is mounted to a system bus of a host server, the host server further including

at least one host processor connected to the system bus, and

a network interface device;

encapsulating the network packet data to create encapsulated network packets for transport on a logical network or decapsulating the network packet data to create decapsulated network packets for delivery to a network location, the encapsulating and decapsulating being executed by processing circuits mounted on the offload processor module and being executed independent of any host processor; and

transporting the encapsulated network packets or the decapsulated network packets out of the offload processor module; wherein

the logical network is overlaid on a physical network.

#### **A. NVIDIA'S DIRECT INFRINGEMENT**

367. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '924 Patent, including at least Claim 9.

368. The NVIDIA Accused Products provide network overlay services (e.g., network overlay services for scaling of data center networks by offloading, accelerating, and isolating the infrastructure services from the host processor).

369. The NVIDIA Accused Products receive network packet data from a data source in an offload processor module (e.g., an NVIDIA BlueField DPU) that is mounted to a system bus (e.g., a PCIe bus) of a host server.

370. The host server of the NVIDIA Accused Products includes at least one host processor connected to the system bus and a network interface device (e.g., a ConnectX DPU).

371. In addition, the NVIDIA Accused Products encapsulate the network packet data to create encapsulated network packets for transport on a logical network and/or decapsulate the network packet data to create decapsulated network packets for delivery to a network location

(e.g., the NVIDIA BlueField DPU provides advanced hardware offloading engines that encapsulate/decapsulate overlay protocols).

372. In the NVIDIA Accused Products, the encapsulating and decapsulating is executed by processing circuits mounted on the offload processor module (e.g., by hardware offload engines of the ConnectX DPU on the NVIDIA BlueField DPU).

373. Further, in the NVIDIA Accused Products, the encapsulating and decapsulating is executed independent of any host processor (e.g., the encapsulation and decapsulation functions of the ConnectX DPU are offloaded, accelerated, and isolated from the CPU).

374. The NVIDIA Accused Products transport the encapsulated network packets and/or the decapsulated network packets out of the offload processor module (e.g., out of the NVIDIA BlueField DPU and over the logical tunnels of the overlay networks).

375. Lastly, in the NVIDIA Accused Products, the logical network is overlaid on a physical network (e.g., using overlay technologies as Virtual eXtensible Local-Area Network (VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), and Generic Network Virtualization Encapsulation (GENEVE)).

## **B. MICROSOFT'S DIRECT INFRINGEMENT**

376. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '924 Patent, including at least Claim 9.

377. The Microsoft Accused Products provide network overlay services (e.g., network overlay services for scaling of data center networks by offloading, accelerating, and isolating the infrastructure services from the host processor).

378. The Microsoft Accused Products receive network packet data from a data source in an offload processor module (e.g., an NVIDIA BlueField DPU) that is mounted to a system bus (e.g., a PCIe bus) of a host server.

379. The host server of the Microsoft Accused Products includes at least one host processor connected to the system bus and a network interface device (e.g., a ConnectX DPU).

380. In addition, the Microsoft Accused Products encapsulate the network packet data to create encapsulated network packets for transport on a logical network and/or decapsulate the network packet data to create decapsulated network packets for delivery to a network location (e.g., the NVIDIA BlueField DPU provides advanced hardware offloading engines that encapsulate/decapsulate overlay protocols).

381. In the Microsoft Accused Products, the encapsulating and decapsulating is executed by processing circuits mounted on the offload processor module (e.g., by hardware offload engines of the ConnectX DPU on the NVIDIA BlueField DPU).

382. Further, in the Microsoft Accused Products, the encapsulating and decapsulating is executed independent of any host processor (e.g., the encapsulation and decapsulation functions of the ConnectX DPU are offloaded, accelerated, and isolated from the CPU).

383. The Microsoft Accused Products transport the encapsulated network packets and/or the decapsulated network packets out of the offload processor module (e.g., out of the NVIDIA BlueField DPU and over the logical tunnels of the overlay networks).

384. Lastly, in the Microsoft Accused Products, the logical network is overlaid on a physical network (e.g., using overlay technologies as Virtual eXtensible Local-Area Network (VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), and Generic Network Virtualization Encapsulation (GENEVE)).



## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

385. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '924 Patent's claims, including at least Claim 9, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '924 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '924 Patent, including at least Claim 9.

386. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '924 Patent, including at least Claim 9, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that would infringe one or more of the claims of the '924 Patent, including at least Claim 9.

387. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA

ConnectX DPUs, in manners that would infringe one or more of the claims of the '924 Patent, including at least Claim 9.

388. NVIDIA has had knowledge of the '924 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA BlueField DPU and ConnectX DPU in server systems would infringe Xockets' Asserted Patents, including the '924 Patent. To the extent that NVIDIA lacked actual knowledge of the '924 Patent or its customers' and/or end-users' actual infringement of the '924 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

389. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '924 Patent, including at least Claim 9, with knowledge of the '924 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '924 Patent.

390. At a minimum, NVIDIA has had actual notice of the '924 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 9 of the '924 Patent by its customers and end-users, including Microsoft.

391. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '924 Patent's claims, including at least Claim 9, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that

incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '924 Patent, including at least Claim 9.

392. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '924 Patent, including at least Claim 9, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

393. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '924 Patent, including at least Claim 9.

#### **B. MICROSOFT'S INDIRECT INFRINGEMENT**

394. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '924 Patent's claims, including at least Claim 9, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '924 Patent, including at least Claim 9.

395. For example, Microsoft sells over 200 Azure products<sup>101</sup> and over 40 Azure cloud solutions<sup>102</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

396. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '924 Patent, including at least Claim 9, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '924 Patent, including at least Claim 9.

397. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '924 Patent, including at least Claim 9.

398. Microsoft has had knowledge of the '924 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '924

---

<sup>101</sup> <https://azure.microsoft.com/en-us/products>.

<sup>102</sup> <https://azure.microsoft.com/en-us/solutions>.

Patent. To the extent that Microsoft lacked actual knowledge of the '924 Patent or its customers' and/or end-users' actual infringement of the '924 Patent, Microsoft took deliberate actions to avoid learning of those facts.

399. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '924 Patent, including at least Claim 19, with knowledge of the '924 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '924 Patent.

400. At a minimum, Microsoft has had actual notice of the '924 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 9 of the '924 Patent by its customers and end-users.

401. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '924 Patent's claims, including at least Claim 9, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '924 Patent, including at least Claim 9.

402. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '924 Patent, including at least Claim 9, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to

perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

403. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '924 Patent, including at least Claim 9.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

404. NVIDIA has had knowledge of the '924 Patent no later than February 10, 2022.

405. Despite knowing of the '924 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '924 Patent.

406. Despite knowing of the '924 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '924 Patent.

407. At a minimum, NVIDIA has had actual notice of the '924 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '924 Patent, including at least Claim 9.

408. At a minimum, NVIDIA has also willfully blinded itself to the '924 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '924 Patent but took deliberate steps to avoid learning of its infringement.

409. Therefore, upon information and belief, NVIDIA's infringement of the '924 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

410. Microsoft has had knowledge of the '924 Patent no later than March 22, 2017.

411. Despite knowing of the '924 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '924 Patent.

412. Despite knowing of the '924 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '924 Patent.

413. At a minimum, Microsoft has had actual notice of the '924 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '924 Patent, including at least Claim 9.

414. At a minimum, Microsoft has also willfully blinded itself to the '924 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '924 Patent but took deliberate steps to avoid learning of its infringement.

415. Therefore, upon information and belief, Microsoft's infringement of the '924 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**COUNT V: INFRINGEMENT OF THE '350 PATENT**

416. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

**I. DIRECT INFRINGEMENT**

417. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '350 Patent's claims, including at least Claim 1, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA BlueField DPUs and ConnectX DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

418. NVIDIA and Microsoft are infringing claims of the '350 Patent, including at least Claim 1, literally and/or pursuant to the doctrine of equivalents.

419. Claim 1 of the '350 Patent is directed to a device, comprising:

- a server that includes a host processor and at least one hardware acceleration (hwa) module physically separate from the host processor and having

- a network interface configured to virtualize functions by redirecting network packets to different addresses within the hwa,

- at least one computing element formed thereon, the at least one computing element including

- processing circuits configured to execute a plurality of processes including at least one virtualized function,

- a scheduler circuit configured to allocate a priority to a processing of packets of one flow over those of another flow by the processing circuits,

- first memory circuits,

- second memory circuits, and

- a data transfer fabric configured to enable data transfers between the processing circuits and the first and second memory circuits; wherein

- the at least one computing element is configured to transfer data to, or receive data from, any of: the processing circuits, the first



memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric.

**A. NVIDIA'S DIRECT INFRINGEMENT**

420. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '350 Patent, including at least Claim 1.

421. The NVIDIA Accused Products are devices that comprise a server (e.g., an NVIDIA GPU-centric server) that includes a host processor (e.g., a host CPU) and at least one hardware acceleration module physically separate from the host processor (e.g., an NVIDIA BlueField DPU with hardware accelerators, which allow the DPU to deliver a broad set of accelerated software-defined networking, storage, security, and management services with the ability to offload, accelerate and isolate data center infrastructure).

422. The server of the NVIDIA Accused Products further includes a network interface (e.g., a ConnectX DPU with Ethernet/InfiniBand interfaces) configured to virtualize functions by redirecting network packets to different addresses within the hardware acceleration module (e.g., with eSwitch Flow Steering/Switching functions).

423. In addition, the server of the NVIDIA Accused Products includes at least one computing element formed thereon (e.g., ARM cores and their associated cache, memory, and scheduler).

424. The computing element of the NVIDIA Accused Products includes processing circuits configured to execute a plurality of processes including at least one virtualized function (e.g., custom accelerations for distributed machine learning and AI applications).

425. The computing element of the NVIDIA Accused Products also includes a scheduler circuit configured to allocate a priority to a processing of packets of one flow over those of another

flow by the processing circuits (e.g., in the ConnectX DPU, which provides advanced packet processing functionalities for streaming data into and out of the ARM cores, cache, and memory controllers).

426. The computing element of the NVIDIA Accused Products further includes first and second memory circuits (e.g., L2/L3 cache and DDR4 memory controllers).

427. The server of the NVIDIA Accused Products also includes a data transfer fabric (e.g., a Cache Coherent Mesh Interconnect) configured to enable data transfers between the processing circuits and the first and second memory circuits.

428. Lastly, in the NVIDIA Accused Products, the computing element is configured to transfer data to, or receive data from, any of the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric.

## **B. MICROSOFT'S DIRECT INFRINGEMENT**

429. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '350 Patent, including at least Claim 1.

430. The Microsoft Accused Products are devices that comprise a server (e.g., an NVIDIA or other server) that includes a host processor (e.g., a host CPU) and at least one hardware acceleration module physically separate from the host processor (e.g., an NVIDIA BlueField DPU with hardware accelerators, which allow the DPU to deliver a broad set of accelerated software-defined networking, storage, security, and management services with the ability to offload, accelerate and isolate data center infrastructure).

431. The server of the Microsoft Accused Products further includes a network interface (e.g., a ConnectX DPU with Ethernet/InfiniBand interfaces) configured to virtualize functions by

redirecting network packets to different addresses within the hardware acceleration module (e.g., with eSwitch Flow Steering/Switching functions).

432. In addition, the server of the Microsoft Accused Products includes at least one computing element formed thereon (e.g., ARM cores and their associated cache, memory, and scheduler).

433. The computing element of the Microsoft Accused Products includes processing circuits configured to execute a plurality of processes including at least one virtualized function (e.g., custom accelerations for distributed machine learning and AI applications).

434. The computing element of the Microsoft Accused Products also includes a scheduler circuit configured to allocate a priority to a processing of packets of one flow over those of another flow by the processing circuits (e.g., in the ConnectX DPU, which provides advanced packet processing functionalities for streaming data into and out of the ARM cores, cache, and memory controllers).

435. The computing element of the Microsoft Accused Products further includes first and second memory circuits (e.g., L2/L3 cache and DDR4 memory controllers).

436. In addition, the server of the Microsoft Accused Products includes a data transfer fabric (e.g., a Cache Coherent Mesh Interconnect) configured to enable data transfers between the processing circuits and the first and second memory circuits.

437. Lastly, in the Microsoft Accused Products, the computing element is configured to transfer data to, or receive data from, any of the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric.

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

438. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '350 Patent's claims, including at least Claim 1, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '350 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '350 Patent, including at least Claim 1.

439. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '350 Patent, including at least Claim 1, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that would infringe one or more of the claims of the '350 Patent, including at least Claim 1.

440. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server systems incorporating NVIDIA BlueField DPUs and NVIDIA

ConnectX DPUs, in manners that would infringe one or more of the claims of the '350 Patent, including at least Claim 1.

441. NVIDIA has had knowledge of the '350 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA BlueField DPU and ConnectX DPU in server systems would infringe Xockets' Asserted Patents, including the '350 Patent. To the extent that NVIDIA lacked actual knowledge of the '350 Patent or its customers' and/or end-users' actual infringement of the '350 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

442. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '350 Patent, including at least Claim 1, with knowledge of the '350 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '350 Patent.

443. At a minimum, NVIDIA has had actual notice of the '350 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '350 Patent by its customers and end-users, including Microsoft.

444. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '350 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that

incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '350 Patent, including at least Claim 1.

445. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '350 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

446. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '350 Patent, including at least Claim 1.

#### **B. MICROSOFT'S INDIRECT INFRINGEMENT**

447. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '350 Patent's claims, including at least Claim 1, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '350 Patent, including at least Claim 1.

448. For example, Microsoft sells over 200 Azure products<sup>103</sup> and over 40 Azure cloud solutions<sup>104</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

449. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '350 Patent, including at least Claim 1, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '350 Patent, including at least Claim 1.

450. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '350 Patent, including at least Claim 1.

451. Microsoft has had knowledge of the '350 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '350

---

<sup>103</sup> <https://azure.microsoft.com/en-us/products>.

<sup>104</sup> <https://azure.microsoft.com/en-us/solutions>.

Patent. To the extent that Microsoft lacked actual knowledge of the '350 Patent or its customers' and/or end-users' actual infringement of the '350 Patent, Microsoft took deliberate actions to avoid learning of those facts.

452. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '350 Patent, including at least Claim 1, with knowledge of the '350 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '350 Patent.

453. At a minimum, Microsoft has had actual notice of the '350 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '350 Patent by its customers and end-users.

454. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '350 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA BlueField DPUs and NVIDIA ConnectX DPUs in manners that infringe the '350 Patent, including at least Claim 1.

455. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA BlueField DPUs and ConnectX DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '350 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to



perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

456. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '350 Patent, including at least Claim 1.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

457. NVIDIA has had knowledge of the '350 Patent no later than February 10, 2022.

458. Despite knowing of the '350 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '350 Patent.

459. Despite knowing of the '350 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '350 Patent.

460. At a minimum, NVIDIA has had actual notice of the '350 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '350 Patent, including at least Claim 1.

461. At a minimum, NVIDIA has also willfully blinded itself to the '350 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '350 Patent but took deliberate steps to avoid learning of its infringement.

462. Therefore, upon information and belief, NVIDIA's infringement of the '350 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

463. Microsoft has had knowledge of the '350 Patent no later than March 22, 2017.

464. Despite knowing of the '350 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '350 Patent.

465. Despite knowing of the '350 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '350 Patent.

466. At a minimum, Microsoft has had actual notice of the '350 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '350 Patent, including at least Claim 1.

467. At a minimum, Microsoft has also willfully blinded itself to the '350 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '350 Patent but took deliberate steps to avoid learning of its infringement.

468. Therefore, upon information and belief, Microsoft's infringement of the '350 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**COUNT VI: INFRINGEMENT OF THE '297 PATENT**

469. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

**I. DIRECT INFRINGEMENT**

470. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '297 Patent's claims, including at least Claim 1 and Claim 7, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA NVLink Switch DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

471. NVIDIA and Microsoft are infringing claims of the '297 Patent, including at least Claim 1 and Claim 7, literally and/or pursuant to the doctrine of equivalents.

472. Claim 1 of the '297 Patent is directed to a system, comprising:

a plurality of first server modules interconnected to one another via  
a communication network, each first server module including

a first switch,

at least one main processor, and

at least one computation module coupled to the main processor  
by a bus, each computation module including

a second switch, and

a plurality of computation elements; wherein

the second switches of the first server modules form a switching  
plane for the ingress and egress of network packets independent of  
any main processors of the first server modules, and

each computation module is insertable into a physical connector of  
the first server module.

473. Claim 7 of the '297 Patent is directed to:

The system of claim 1, wherein the second switch is a virtual switch  
comprising computation elements on the computation module.

**A. NVIDIA'S DIRECT INFRINGEMENT**

474. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '297 Patent, including at least Claim 1 and Claim 7.

475. The NVIDIA Accused Products are systems comprising a plurality of first server modules (e.g., NVIDIA GPU-centric server modules) that are interconnected to one another by a communication network (e.g., NVIDIA's Quantum InfiniBand network).

476. Each first server module in the NVIDIA Accused Products includes a first switch (e.g., an NVIDIA ConnectX DPU and/or NVIDIA BlueField DPU), at least one main processor (e.g., an NVIDIA Superchip), and at least one computation module coupled to the main processor by a bus (e.g., an NVIDIA NVLink Switch DPU coupled to the Superchip by an NVLink Cable Cartridge bus).

477. In the NVIDIA Accused Products, each computation module includes a second switch and a plurality of computation elements (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs).

478. The second switches of the first server modules in the NVIDIA Accused Products form a switching plane for the ingress and egress of network packets independent of any main processors of the first server modules (e.g., a switching plane formed by the NVLink Switch DPUs in an NVLink domain independent of the NVIDIA Superchips).

479. Each computation module in the NVIDIA Accused Products is insertable into a physical connector of the first server module (e.g., each NVLink Switch DPU is insertable into a physical connector of the server module).

480. Lastly, the second switch in the NVIDIA Accused Products is a virtual switch comprising computation elements on the computation module.

**B. MICROSOFT'S DIRECT INFRINGEMENT**

481. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '297 Patent, including at least Claim 1 and Claim 7.

482. The Microsoft Accused Products are systems comprising a plurality of first server modules (e.g., NVIDIA or other server modules) that are interconnected to one another by a communication network (e.g., NVIDIA's Quantum InfiniBand network).

483. Each first server module in the Microsoft Accused Products includes a first switch (e.g., an NVIDIA ConnectX DPU and/or NVIDIA BlueField DPU), at least one main processor (e.g., an NVIDIA Superchip), and at least one computation module coupled to the main processor by a bus (e.g., an NVIDIA NVLink Switch DPU coupled to the Superchip by an NVLink Cable Cartridge bus).

484. In the Microsoft Accused Products, each computation module includes a second switch and a plurality of computation elements (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs).

485. The second switches of the first server modules in the Microsoft Accused Products form a switching plane for the ingress and egress of network packets independent of any main processors of the first server modules (e.g., a switching plane formed by the NVLink Switch DPUs in an NVLink domain independent of the NVIDIA Superchips).

486. Each computation module in the Microsoft Accused Products is insertable into a physical connector of the first server module (e.g., each NVLink Switch DPU is insertable into a physical connector of the server module).

487. Lastly, the second switch in the Microsoft Accused Products is a virtual switch comprising computation elements on the computation module.

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

488. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '297 Patent's claims, including at least Claim 1 and Claim 7, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '297 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '297 Patent, including at least Claim 1 and Claim 7.

489. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '297 Patent, including at least Claim 1 and Claim 7, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA NVLink Switch DPUs in manners that would infringe one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

490. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA NVLink Switch DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server

systems incorporating NVIDIA NVLink Switch DPUs, in manners that would infringe one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

491. NVIDIA has actual knowledge of the '297 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA NVLink Switch DPUs in server systems would infringe Xockets' Asserted Patents, including the '297 Patent. To the extent that NVIDIA lacked actual knowledge of the '297 Patent or its customers' and/or end-users' actual infringement of the '297 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

492. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7, with knowledge of the '297 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '297 Patent.

493. At a minimum, NVIDIA has had actual notice of the '297 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 and Claim 7 of the '297 Patent by its customers and end-users, including Microsoft.

494. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '297 Patent's claims, including at least Claim 1 and Claim 7, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused

Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '297 Patent, including at least Claim 1 and Claim 7.

495. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '297 Patent, including at least Claim 1 and Claim 7, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

496. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

## **B. MICROSOFT'S INDIRECT INFRINGEMENT**

497. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '297 Patent's claims, including at least Claim 1 and Claim 7, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '297 Patent, including at least Claim 1 and Claim 7.



498. For example, Microsoft sells over 200 Azure products<sup>105</sup> and over 40 Azure cloud solutions<sup>106</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

499. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '297 Patent, including at least Claim 1 and Claim 7, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

500. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

501. Microsoft has had knowledge of the '297 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '297

---

<sup>105</sup> <https://azure.microsoft.com/en-us/products>.

<sup>106</sup> <https://azure.microsoft.com/en-us/solutions>.

Patent. To the extent that Microsoft lacked actual knowledge of the '297 Patent or its customers' and/or end-users' actual infringement of the '297 Patent, Microsoft took deliberate actions to avoid learning of those facts.

502. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7, with knowledge of the '297 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '297 Patent.

503. At a minimum, Microsoft has had actual notice of the '297 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 and Claim 7 of the '297 Patent by its customers and end-users.

504. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '297 Patent's claims, including at least Claim 1 and Claim 7, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '297 Patent, including at least Claim 1 and Claim 7.

505. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '297 Patent, including at least Claim 1 and Claim 7, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform

the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

506. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '297 Patent, including at least Claim 1 and Claim 7.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

507. NVIDIA has had knowledge of the '297 Patent no later than February 10, 2022.

508. Despite knowing of the '297 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '297 Patent.

509. Despite knowing of the '297 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '297 Patent.

510. At a minimum, NVIDIA has had actual notice of the '297 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '297 Patent, including at least Claim 1 and Claim 7.

511. At a minimum, NVIDIA has also willfully blinded itself to the '297 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '297 Patent but took deliberate steps to avoid learning of its infringement.

512. Therefore, upon information and belief, NVIDIA's infringement of the '297 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

513. Microsoft has had knowledge of the '297 Patent no later than March 22, 2017.

514. Despite knowing of the '297 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '297 Patent.

515. Despite knowing of the '297 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '297 Patent.

516. At a minimum, Microsoft has had actual notice of the '297 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '297 Patent, including at least Claim 1 and Claim 7.

517. At a minimum, Microsoft has also willfully blinded itself to the '297 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '297 Patent but took deliberate steps to avoid learning of its infringement.

518. Therefore, upon information and belief, Microsoft's infringement of the '297 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**COUNT VII: INFRINGEMENT OF THE '161 PATENT**

519. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

## **I. DIRECT INFRINGEMENT**

520. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '161 Patent's claims, including at least Claim 1, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA NVLink Switch DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

521. NVIDIA and Microsoft are infringing claims of the '161 Patent, including at least Claim 1, literally and/or pursuant to the doctrine of equivalents.

522. Claim 1 of the '161 Patent is directed to a rack server system for a packet processing, comprising:

- a plurality of servers mountable in a rack;

- a top of rack (TOR) unit having connections to each of the servers;

- a plurality of offload processor modules, each offload processor module having at least one input-output (IO) port and multiple offload processors, including at least a first offload processor module connected directly to a second offload processor module through their respective IO ports, the offload processor modules are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module; and

- a memory controller configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

### **A. NVIDIA'S DIRECT INFRINGEMENT**

523. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '161 Patent, including at least Claim 1.

524. The NVIDIA Accused Products are rack server systems for a packet processing, which include a plurality of servers (e.g., NVIDIA GPU-centric servers) mountable in a server rack.

525. The rack server systems of the NVIDIA Accused Products also include a top of rack unit having connections to each of the servers (e.g., a top of rack switch in NVIDIA's Quantum InfiniBand network).

526. The rack server systems of the NVIDIA Accused Products further include a plurality of offload processor modules (e.g., NVIDIA NVLink Switch DPUs), each having at least one input-output ("IO") port (e.g., NVLink ports) and multiple offload processors (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs).

527. The plurality of offload processor modules in the NVIDIA Accused Products include at least a first offload processor module connected directly to a second offload processor module through their respective IO ports (e.g., an NVLink Switch DPU connected to another NVLink Switch DPU through their respective IO ports to form a non-blocking switching fabric).

528. In addition, the offload processor modules in the NVIDIA Accused Products are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module (e.g., from the NVLink ports on the NVLink Switch DPU).

529. Lastly, the rack server systems of the NVIDIA Accused Products include a memory controller (e.g., an LPDDR5 and/or HBM memory controller of the NVIDIA Superchips) configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

**B. MICROSOFT'S DIRECT INFRINGEMENT**

530. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '161 Patent, including at least Claim 1.

531. The Microsoft Accused Products are rack server systems for a packet processing, which include a plurality of servers (e.g., NVIDIA or other servers) mountable in a server rack.

532. The rack server systems of the Microsoft Accused Products also include a top of rack unit having connections to each of the servers (e.g., a top of rack switch in NVIDIA's Quantum InfiniBand network).

533. The rack server systems of the Microsoft Accused Products further include a plurality of offload processor modules (e.g., NVIDIA NVLink Switch DPUs), each having at least one input-output ("IO") port (e.g., NVLink ports) and multiple offload processors (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs).

534. The plurality of offload processor modules in the Microsoft Accused Products include at least a first offload processor module connected directly to a second offload processor module through their respective IO ports (e.g., an NVLink Switch DPU connected to another NVLink Switch DPU through their respective IO ports to form a non-blocking switching fabric).

535. In addition, the offload processor modules in the Microsoft Accused Products are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module (e.g., from the NVLink ports on the NVLink Switch DPU).

536. Lastly, the rack server systems of the Microsoft Accused Products include a memory controller (e.g., an LPDDR5 and/or HBM memory controller of the NVIDIA Superchips)

configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

537. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '161 Patent's claims, including at least Claim 1, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '161 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '161 Patent, including at least Claim 1.

538. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '161 Patent, including at least Claim 1, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA NVLink Switch DPUs in manners that would infringe one or more of the claims of the '161 Patent, including at least Claim 1.

539. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA NVLink Switch DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server



systems incorporating NVIDIA NVLink Switch DPUs, in manners that would infringe one or more of the claims of the '161 Patent, including at least Claim 1.

540. NVIDIA has actual knowledge of the '161 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA NVLink Switch DPUs in server systems would infringe Xockets' Asserted Patents, including the '161 Patent. To the extent that NVIDIA lacked actual knowledge of the '161 Patent or its customers' and/or end-users' actual infringement of the '161 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

541. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '161 Patent, including at least Claim 1, with knowledge of the '161 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '092 Patent.

542. At a minimum, NVIDIA has had actual notice of the '161 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '161 Patent by its customers and end-users, including Microsoft.

543. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '161 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that

incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '161 Patent, including at least Claim 1.

544. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '161 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

545. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '161 Patent, including at least Claim 1.

#### **B. MICROSOFT'S INDIRECT INFRINGEMENT**

546. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '161 Patent's claims, including at least Claim 1, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '161 Patent, including at least Claim 1.

547. For example, Microsoft sells over 200 Azure products<sup>107</sup> and over 40 Azure cloud solutions<sup>108</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

548. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '161 Patent, including at least Claim 1, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '161 Patent, including at least Claim 1.

549. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '161 Patent, including at least Claim 1.

550. Microsoft has had knowledge of the '161 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '161

---

<sup>107</sup> <https://azure.microsoft.com/en-us/products>.

<sup>108</sup> <https://azure.microsoft.com/en-us/solutions>.

Patent. To the extent that Microsoft lacked actual knowledge of the '161 Patent or its customers' and/or end-users' actual infringement of the '161 Patent, Microsoft took deliberate actions to avoid learning of those facts.

551. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '161 Patent, including at least Claim 1, with knowledge of the '161 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '161 Patent.

552. At a minimum, Microsoft has had actual notice of the '161 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '161 Patent by its customers and end-users.

553. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '161 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '161 Patent, including at least Claim 1.

554. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '161 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed

functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

555. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '161 Patent, including at least Claim 1.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

556. NVIDIA has had knowledge of the '161 Patent no later than February 10, 2022.

557. Despite knowing of the '161 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '161 Patent.

558. Despite knowing of the '161 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '161 Patent.

559. At a minimum, NVIDIA has had actual notice of the '161 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '161 Patent, including at least Claim 1.

560. At a minimum, NVIDIA has also willfully blinded itself to the '161 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '161 Patent but took deliberate steps to avoid learning of its infringement.

561. Therefore, upon information and belief, NVIDIA's infringement of the '161 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

562. Microsoft has had knowledge of the '161 Patent no later than March 22, 2017.

563. Despite knowing of the '161 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '161 Patent.

564. Despite knowing of the '161 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '161 Patent.

565. At a minimum, Microsoft has had actual notice of the '161 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '161 Patent, including at least Claim 1.

566. At a minimum, Microsoft has also willfully blinded itself to the '161 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '161 Patent but took deliberate steps to avoid learning of its infringement.

567. Therefore, upon information and belief, Microsoft's infringement of the '161 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**COUNT VIII: INFRINGEMENT OF THE '092 PATENT**

568. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

**I. DIRECT INFRINGEMENT**

569. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '092 Patent's claims, including at least Claim 1, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA NVLink Switch DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

570. NVIDIA and Microsoft are infringing claims of the '092 Patent, including at least Claim 1, literally and/or pursuant to the doctrine of equivalents.

571. Claim 1 of the '092 Patent is directed to a distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data, the architecture comprising:

- a plurality of servers, including first servers that each include

- at least one central processing unit (CPU), and

- at least one offload processing module coupled to the at least one CPU by a bus, each offload processing module including a plurality of computation elements, the computation elements configured to

- operate as a virtual switch, and

- execute the second computing operations on first processed data to generate second processed data; wherein

- the virtual switches form a switch fabric for exchanging data between the offload processing modules,

- the first computing operations generate the first processed data and are not executed by the offload processing modules, and

- the second computing operations are executed on a plurality of the offload processing modules in parallel.

**A. NVIDIA'S DIRECT INFRINGEMENT**

572. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '092 Patent, including at least Claim 1.

573. The NVIDIA Accused Products are a distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data rack server systems for a packet processing (e.g., NVIDIA's GPU-centric server systems for distributed computing).

574. The architecture of the NVIDIA Accused Products includes a plurality of servers (e.g., NVIDIA GPU-centric servers) that each include at least one CPU (e.g., in the NVIDIA Superchips) and at least one offload processing module (e.g., NVIDIA NVLink Switch DPUs) coupled to the CPU by a bus (e.g., an NVLink Cable Cartridge bus).

575. Each offload processing module in the NVIDIA Accused Products includes a plurality of computation elements (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs) that are configured to operate as a virtual switch.

576. In the architecture of the NVIDIA Accused Products, the virtual switches form a switch fabric for exchanging data between the offload processing modules (e.g., form a switching plane for exchanging data between the NVLink Switch DPUs).

577. In addition, in the architecture of the NVIDIA Accused Products, first computing operations generate the first processed data and are not executed by the offload processing modules (e.g., in machine learning/AI training processes involving all-reduce operations, GPUs perform arithmetic operations to generate local gradients, and the arithmetic operations are not executed by the NVLink Switch DPUs).



578. The computation elements in each offload processing module in the NVIDIA Accused Products are configured to execute second computing operations on the first processed data to generate second processed data (e.g., the GPUs send the processed data to the hardware accelerator engines in the NVLink Switch DPUs, which perform further computing operations, including reductions, to generate results).

579. Lastly, in the architecture of the NVIDIA Accused Products, the second computing operations are executed on a plurality of the offload processing modules in parallel (e.g., the NVLink Switch DPUs perform the calculations in parallel).

## **B. MICROSOFT'S DIRECT INFRINGEMENT**

580. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '092 Patent, including at least Claim 1.

581. The Microsoft Accused Products are a distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data rack server systems for a packet processing (e.g., Microsoft's Azure Cloud deployed on NVIDIA GPU-centric server systems with an AI supercomputer architecture for distributed computing).

582. The architecture of the Microsoft Accused Products includes a plurality of servers (e.g., NVIDIA or other servers) that each include at least one CPU (e.g., in the NVIDIA Superchips) and at least one offload processing module (e.g., NVIDIA NVLink Switch DPUs) coupled to the CPU by a bus (e.g., an NVLink Cable Cartridge bus).

583. Each offload processing module in the Microsoft Accused Products includes a plurality of computation elements (e.g., hardware acceleration engines on NVIDIA NVLink Switch DPUs) that are configured to operate as a virtual switch.

584. In the architecture of the Microsoft Accused Products, the virtual switches form a switch fabric for exchanging data between the offload processing modules (e.g., form a switching plane for exchanging data between the NVLink Switch DPUs).

585. In addition, in the architecture of the Microsoft Accused Products, first computing operations generate the first processed data and are not executed by the offload processing modules (e.g., in machine learning/AI training processes involving all-reduce operations, GPUs perform arithmetic operations to generate local gradients, and the arithmetic operations are not executed by the NVLink Switch DPUs).

586. The computation elements in each offload processing module in the Microsoft Accused Products are configured to execute second computing operations on the first processed data to generate second processed data (e.g., the GPUs send the processed data to the hardware accelerator engines in the NVLink Switch DPUs, which perform further computing operations, including reductions, to generate results).

587. Lastly, in the architecture of the Microsoft Accused Products, the second computing operations are executed on a plurality of the offload processing modules in parallel (e.g., the NVLink Switch DPUs perform the calculations in parallel).

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

588. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '092 Patent's claims, including at least Claim 1, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '092 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use

of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '092 Patent, including at least Claim 1.

589. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '092 Patent, including at least Claim 1, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA NVLink Switch DPUs in manners that would infringe one or more of the claims of the '092 Patent, including at least Claim 1.

590. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA NVLink Switch DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server systems incorporating NVIDIA NVLink Switch DPUs, in manners that would infringe one or more of the claims of the '092 Patent, including at least Claim 1.

591. NVIDIA has actual knowledge of the '092 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA NVLink Switch DPUs in server systems would infringe Xockets' Asserted Patents, including the '092 Patent. To the extent that NVIDIA lacked actual knowledge

of the '092 Patent or its customers' and/or end-users' actual infringement of the '092 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

592. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '092 Patent, including at least Claim 1, with knowledge of the '092 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '092 Patent.

593. At a minimum, NVIDIA has had actual notice of the '092 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '092 Patent by its customers and end-users, including Microsoft.

594. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '092 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '092 Patent, including at least Claim 1.

595. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '092 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to

perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

596. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '092 Patent, including at least Claim 1.

## **B. MICROSOFT'S INDIRECT INFRINGEMENT**

597. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '092 Patent's claims, including at least Claim 1, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '092 Patent, including at least Claim 1.

598. For example, Microsoft sells over 200 Azure products<sup>109</sup> and over 40 Azure cloud solutions<sup>110</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

599. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '092 Patent, including at least Claim 1, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '092 Patent, including at least Claim 1.

---

<sup>109</sup> <https://azure.microsoft.com/en-us/products>.

<sup>110</sup> <https://azure.microsoft.com/en-us/solutions>.

600. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '092 Patent, including at least Claim 1.

601. Microsoft has had knowledge of the '092 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '092 Patent. To the extent that Microsoft lacked actual knowledge of the '092 Patent or its customers' and/or end-users' actual infringement of the '092 Patent, Microsoft took deliberate actions to avoid learning of those facts.

602. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '092 Patent, including at least Claim 1, with knowledge of the '092 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '092 Patent.

603. At a minimum, Microsoft has had actual notice of the '092 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 1 of the '092 Patent by its customers and end-users.

604. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '092 Patent's claims, including at least Claim 1, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '092 Patent, including at least Claim 1.

605. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '092 Patent, including at least Claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

606. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '092 Patent, including at least Claim 1.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

607. NVIDIA has had knowledge of the '092 Patent no later than February 10, 2022.

608. Despite knowing of the '092 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '092 Patent.

609. Despite knowing of the '092 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '092 Patent.

610. At a minimum, NVIDIA has had actual notice of the '092 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '092 Patent, including at least Claim 1.

611. At a minimum, NVIDIA has also willfully blinded itself to the '092 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '092 Patent but took deliberate steps to avoid learning of its infringement.

612. Therefore, upon information and belief, NVIDIA's infringement of the '092 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

## **B. MICROSOFT'S WILLFUL INFRINGEMENT**

613. Microsoft has had knowledge of the '092 Patent no later than March 22, 2017.

614. Despite knowing of the '092 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '092 Patent.

615. Despite knowing of the '092 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '092 Patent.

616. At a minimum, Microsoft has had actual notice of the '092 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '092 Patent, including at least Claim 1.



617. At a minimum, Microsoft has also willfully blinded itself to the '092 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '092 Patent but took deliberate steps to avoid learning of its infringement.

618. Therefore, upon information and belief, Microsoft's infringement of the '092 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

#### **COUNT IX: INFRINGEMENT OF THE '640 PATENT**

619. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

##### **I. DIRECT INFRINGEMENT**

620. In violation of 35 U.S.C. § 271(a), NVIDIA and Microsoft are and have been directly infringing one or more of the '640 Patent's claims, including at least Claim 9, by making, using, selling, and/or offering for sale in the United States, and/or importing into the United States, without authority, server system products and services, including but not limited to those utilizing the NVIDIA NVLink Switch DPUs, including without limitation the NVIDIA Accused Products and the Microsoft Accused Products, as described above.

621. NVIDIA and Microsoft are infringing claims of the '640 Patent, including at least Claim 9, literally and/or pursuant to the doctrine of equivalents.

622. Claim 9 of the '640 Patent is directed to a rack server system for a map/reduce data processing, comprising:

a plurality of servers arranged in a rack,

a plurality of offload processor modules supported on at least two of the servers, each offload processor module having an input-output (IO) port and multiple offload processors, a first offload processor module configured to execute map steps of the map/reduce data processing, and being connected directly to a second offload processor through their respective IO ports to define a midplane switch, and

a top of rack (TOR) unit connected to each of the servers that does not transfer map/reduce data, wherein

a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module.

**A. NVIDIA'S DIRECT INFRINGEMENT**

623. As to NVIDIA, at least the NVIDIA Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '640 Patent, including at least Claim 9.

624. The NVIDIA Accused Products are rack server systems for a map/reduce data processing, which include a plurality of servers (e.g., NVIDIA GPU-centric servers that provide an accelerated platform for machine learning and AI collective operations) arranged in a server rack.

625. The rack server systems of the NVIDIA Accused Products also include a plurality of offload processor modules (e.g., NVIDIA NVLink Switch DPUs) supported on at least two of the servers.

626. Each offload processor module in the rack system of the NVIDIA Accused Products has an input-output ("IO") port (e.g., NVLink ports) and multiple offload processors (e.g., hardware accelerator engines on NVIDIA NVLink Switch DPUs).

627. Of the plurality of offload processor modules in the NVIDIA Accused Products, a first offload processor module is connected directly to a second offload processor through their

respective IO ports to define a midplane switch (e.g., an NVLink Switch DPU connected to another NVLink Switch DPU through the IO ports to define a midplane switch).

628. In addition, the rack server system of the NVIDIA Accused Products includes a top of rack unit connected to each of the servers (e.g., a top of rack switch in NVIDIA's Quantum InfiniBand network) that does not transfer map/reduce data.

629. Further, the first offload processor module of the NVIDIA Accused Products is configured to execute map steps of the map/reduce data processing, and a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module (e.g., a first NVLink Switch DPU assigns messages to destination queues of a second NVLink Switch DPU, and the second NVLink Switch DPU performs reduction operations on the provided messages).

## **B. MICROSOFT'S DIRECT INFRINGEMENT**

630. As to Microsoft, at least the Microsoft Accused Products, as defined above, comprise hardware and software components that together practice every element of one or more claims of the '640 Patent, including at least Claim 9.

631. The Microsoft Accused Products are rack server systems for a map/reduce data processing, which include a plurality of servers (e.g., NVIDIA or other servers that provide an accelerated platform for machine learning and AI collective operations) arranged in a server rack.

632. The rack server systems of the Microsoft Accused Products also include a plurality of offload processor modules (e.g., NVIDIA NVLink Switch DPUs) supported on at least two of the servers.

633. Each offload processor module in the rack system of the Microsoft Accused Products has an input-output ("IO") port (e.g., NVLink ports) and multiple offload processors (e.g., hardware accelerator engines on NVIDIA NVLink Switch DPUs).

634. Of the plurality of offload processor modules in the Microsoft Accused Products, a first offload processor module is connected directly to a second offload processor through their respective IO ports to define a midplane switch (e.g., an NVLink Switch DPU connected to another NVLink Switch DPU through the IO ports to define a midplane switch).

635. In addition, the rack server system of the Microsoft Accused Products includes a top of rack unit connected to each of the servers (e.g., a top of rack switch in NVIDIA's Quantum InfiniBand network) that does not transfer map/reduce data.

636. Further, the first offload processor module of the Microsoft Accused Products is configured to execute map steps of the map/reduce data processing, and a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module (e.g., a first NVLink Switch DPU assigns messages to destination queues of a second NVLink Switch DPU, and the second NVLink Switch DPU performs reduction operations on the provided messages).

## **II. INDIRECT INFRINGEMENT**

### **A. NVIDIA'S INDIRECT INFRINGEMENT**

637. In violation of 35 U.S.C. § 271(b), NVIDIA is and has been infringing one or more of the '640 Patent's claims, including at least Claim 9, indirectly by inducing others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States, to use the NVIDIA Accused Products and/or to make and use other server systems that infringe the '640 Patent. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '640 Patent, including at least Claim 9.

638. Upon information and belief, NVIDIA supplies hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '640 Patent, including at least Claim 9, to induce third parties, including for example NVIDIA's customers and/or end-users, to use the NVIDIA Accused Products and/or make and use other server systems incorporating NVIDIA NVLink Switch DPUs in manners that would infringe one or more of the claims of the '640 Patent, including at least Claim 9.

639. Upon information and belief, NVIDIA furnishes instructive materials, technical support, and information concerning the operation and use of the NVIDIA Accused Products (including the NVIDIA NVLink Switch DPUs) and markets and advertises such products on its website, in videos, at conferences, and elsewhere to induce third parties, including NVIDIA's customers and/or end-users to use the NVIDIA Accused Products, or to make and use other server systems incorporating NVIDIA NVLink Switch DPUs, in manners that would infringe one or more of the claims of the '640 Patent, including at least Claim 9.

640. NVIDIA has actual knowledge of the '640 Patent since at least February 10, 2022. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, NVIDIA subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of the NVIDIA NVLink Switch DPUs in server systems would infringe Xockets' Asserted Patents, including the '640 Patent. To the extent that NVIDIA lacked actual knowledge of the '640 Patent or its customers' and/or end-users' actual infringement of the '640 Patent, NVIDIA took deliberate actions to avoid learning of those facts.

641. Therefore, NVIDIA has induced infringement by others of one or more of the claims of the '640 Patent, including at least Claim 9, with knowledge of the '640 Patent and with

the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '640 Patent.

642. At a minimum, NVIDIA has had actual notice of the '640 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 9 of the '640 Patent by its customers and end-users, including Microsoft.

643. In violation of 35 U.S.C. § 271(c), NVIDIA is and has been infringing one or more of the '640 Patent's claims, including at least Claim 9, indirectly by contributing to the direct infringement committed by others, such as NVIDIA's customers and end-users, in this District and elsewhere in the United States. For example, NVIDIA's customers and/or end-users (including, for example, Microsoft) directly infringe via their use of the NVIDIA Accused Products and/or their manufacture and use of other server systems (such as the Microsoft Accused Products) that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '640 Patent, including at least Claim 9.

644. NVIDIA makes and sells hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '640 Patent, including at least Claim 9, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

645. Therefore, NVIDIA has contributed to the infringement by others of one or more of the claims of the '640 Patent, including at least Claim 9.

**B. MICROSOFT'S INDIRECT INFRINGEMENT**

646. In violation of 35 U.S.C. § 271(b), Microsoft is and has been infringing one or more of the '640 Patent's claims, including at least Claim 9, indirectly by inducing others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States, to use the Microsoft Accused Products. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other features, the NVIDIA NVLink Switch DPUs in manners that infringe the '640 Patent, including at least Claim 9.

647. For example, Microsoft sells over 200 Azure products<sup>111</sup> and over 40 Azure cloud solutions<sup>112</sup> (including products and services relating to AI, machine learning, and high-performance computing) for use by Microsoft's customers and/or end-users.

648. Upon information and belief, Microsoft provides Azure products and services via hardware, firmware, and/or software, including software drivers, that are especially made or especially adapted to practice the inventions claimed in the '640 Patent, including at least Claim 9, to induce third parties, including for example Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '640 Patent, including at least Claim 9.

649. Upon information and belief, Microsoft furnishes instructive materials, technical support, and information concerning the operation and use of the Microsoft Accused Products (including use of Microsoft's Azure products and services) and markets and advertises such products and services on its website, in videos, at conferences, and elsewhere to induce third

---

<sup>111</sup> <https://azure.microsoft.com/en-us/products>.

<sup>112</sup> <https://azure.microsoft.com/en-us/solutions>.

parties, including Microsoft's Azure customers and/or end-users, to use the Microsoft Accused Products in manners that would infringe one or more of the claims of the '640 Patent, including at least Claim 9.

650. Microsoft has had knowledge of the '640 Patent since at least March 22, 2017. Especially in light of its actual knowledge of Xockets and Xockets' patent portfolio, upon information and belief, Microsoft subjectively believed there was a high probability that the Asserted Patents implicated its DPU-enabled server systems and that its customers' and/or end-users' use of these server systems would infringe Xockets' Asserted Patents, including the '640 Patent. To the extent that Microsoft lacked actual knowledge of the '640 Patent or its customers' and/or end-users' actual infringement of the '640 Patent, Microsoft took deliberate actions to avoid learning of those facts.

651. Therefore, Microsoft has induced infringement by others of one or more of the claims of the '640 Patent, including at least Claim 9, with knowledge of the '640 Patent and with the specific intent, or willful blindness, that the induced acts constitute direct infringement of the '640 Patent.

652. At a minimum, Microsoft has had actual notice of the '640 Patent since the filing of this Complaint, yet continues to induce infringement of at least Claim 9 of the '640 Patent by its customers and end-users.

653. In violation of 35 U.S.C. § 271(c), Microsoft is and has been infringing one or more of the '640 Patent's claims, including at least Claim 9, indirectly by contributing to the direct infringement committed by others, such as Microsoft's customers and end-users, in this District and elsewhere in the United States. For example, Microsoft's Azure customers and/or end-users directly infringe via their use of the Microsoft Accused Products that incorporate, among other



features, the NVIDIA NVLink Switch DPUs in manners that infringe the '640 Patent, including at least Claim 9.

654. Microsoft sells at least its Azure products and services on Microsoft Accused Products that include hardware and/or software components (e.g., its NVIDIA NVLink Switch DPUs, along with their software, firmware, and drivers) especially made or especially adapted to practice the invention claimed in the '640 Patent, including at least Claim 9, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

655. Therefore, Microsoft has contributed to the infringement by others of one or more of the claims of the '640 Patent, including at least Claim 9.

### **III. WILLFUL INFRINGEMENT**

#### **A. NVIDIA'S WILLFUL INFRINGEMENT**

656. NVIDIA has had knowledge of the '640 Patent no later than February 10, 2022.

657. Despite knowing of the '640 Patent since at least February 10, 2022, upon information and belief, NVIDIA has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '640 Patent.

658. Despite knowing of the '640 Patent since at least February 10, 2022, NVIDIA has continued to infringe one or more claims of the '640 Patent.

659. At a minimum, NVIDIA has had actual notice of the '640 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '640 Patent, including at least Claim 9.

660. At a minimum, NVIDIA has also willfully blinded itself to the '640 Patent. On information and belief, NVIDIA subjectively believed with a high probability that its NVIDIA Accused Products infringed the '640 Patent but took deliberate steps to avoid learning of its infringement.

661. Therefore, upon information and belief, NVIDIA's infringement of the '640 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

**B. MICROSOFT'S WILLFUL INFRINGEMENT**

662. Microsoft has had knowledge of the '640 Patent no later than March 22, 2017.

663. Despite knowing of the '640 Patent since at least March 22, 2017, upon information and belief, Microsoft has never undertaken any serious investigation to form a good faith belief as to non-infringement or invalidity of the '640 Patent.

664. Despite knowing of the '640 Patent since at least March 22, 2017, Microsoft has continued to infringe one or more claims of the '640 Patent.

665. At a minimum, Microsoft has had actual notice of the '640 Patent, and its infringement thereof, at least as of the filing of this Complaint, yet continues to infringe the '640 Patent, including at least Claim 9.

666. At a minimum, Microsoft has also willfully blinded itself to the '640 Patent. On information and belief, Microsoft subjectively believed with a high probability that its Microsoft Accused Products infringed the '640 Patent but took deliberate steps to avoid learning of its infringement.

667. Therefore, upon information and belief, Microsoft's infringement of the '640 Patent has been and continues to be willful, wanton, malicious, bad-faith, deliberate, consciously wrongful, flagrant, or characteristic of a pirate, entitling Xockets to increased damages pursuant to 35 U.S.C. § 284 and to attorneys' fees and costs incurred in prosecuting this action pursuant to 35 U.S.C. § 285.

### **INJUNCTIVE RELIEF**

668. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

669. Through this complaint, Xockets seeks to enjoin all NVIDIA Accused Products and all Microsoft's use of the Accused Products as a remedy for their willful and unlawful behavior including enjoining the release of NVIDIA's and Microsoft's new Blackwell GPU-enabled server computer systems.

670. Xockets seeks to uphold its Constitutional promise of exclusive rights to its Patents as protected in the United States patent laws. It has only licensed its intellectual property rights on an exclusive basis, including an exclusive license to make and sell Xockets' StreamSwitch.

671. As explained herein above, NVIDIA acknowledges the immense value attributable to the groundbreaking DPU architecture and its significance to the nascent "AI factory" market.

672. NVIDIA continues to expand the scope of its widespread and willful infringement, announcing its plans to release new Blackwell GPU-enabled based server computer systems with DPUs in late 2024 or early 2025.<sup>113</sup> As NVIDIA explains: "Companies and countries are partnering with NVIDIA to shift *the trillion-dollar traditional data centers* to accelerated

---

<sup>113</sup> See <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>; <https://www.fierceelectronics.com/ai/nvidia-expects-billions-blackwell-sales-q4-after-fix>.

computing and build a new type of data center — AI factories — *to produce a new commodity*: artificial intelligence. . . . From server, networking and infrastructure manufacturers to software developers, *the whole industry is gearing up for Blackwell* to accelerate AI-powered innovation for every field.”<sup>114</sup>

673. NVIDIA’s widespread and unauthorized infringement casts a cloud over Xockets’ patents, resulting in a material diminution in their market value and Xockets’ bargaining power. NVIDIA thus usurped unquantifiable value from Xockets’ patented technology and related intellectual property.

674. As a result of NVIDIA’s unauthorized infringement, Xockets has struggled to attract investors, effectively foreclosing Xockets from capitalizing on years of research and development.

675. The release of NVIDIA’s new Blackwell architecture based systems will cause additional irreparable harm to Xockets through further eradication of Xockets’ footprint in this incredibly valuable, emerging technology market; devaluation of Xockets’ patent rights by NVIDIA’s unauthorized and extensive use of Xockets’ properly-granted exclusive rights; and destruction of Xockets’ business opportunities.

676. Not only has Xockets been denied the ability to meaningfully pursue business opportunities related to its patents and patented technology, but Xockets’ only option to obtain *any* compensation for NVIDIA’s extensive and unauthorized use of the patented technology was to pursue litigation.

677. The transaction costs of enforcing patent rights are substantial when combined with large asymmetries in bargaining power between market-controlling entities like NVIDIA and

---

<sup>114</sup> <https://nvidianews.nvidia.com/news/computer-industry-ai-factories-data-centers>.

small business inventors like Xockets. Such costs include substantial litigation costs—in the millions of dollars. The AIPLA reported the *median* litigation costs for patent infringement in 2019 was approximately \$4 million, excluding any parallel challenges to patentability before the Patent Office.<sup>115</sup> Rational investors take into account these costs, and other risks associated with patent enforcement, which discourage investments in innovation.

678. Despite having received Xockets’ notice of infringement, NVIDIA declined to engage in good faith business discussions with Xockets. Rather than pay fair value for Xockets’ exclusive patent rights before incorporating the patented technology in the Accused Products, NVIDIA (leveraging its monopolistic market position and vast resources) chose to employ “an ‘infringe now, pay later’ strategy.”<sup>116</sup>

679. NVIDIA then colluded with other market participants, forming a cartel that engaged RPX to enter into negotiations on their collective behalf. In doing so, NVIDIA unfairly and unlawfully exerted its market influence to further drive down the market value of Xockets’ patent rights and interfered with Xockets’ ability to have good faith business discussions with other market participants.

680. NVIDIA’s refusal to engage in good faith negotiations for the exclusive rights to Xockets’ patents, while continuing to infringe Xockets’ patents—despite communications between the parties in which Xockets offered and demonstrated its willingness to engage in

---

<sup>115</sup> <https://ipwatchdog.com/wp-content/uploads/2021/08/AIPLA-Report-of-the-Economic-Survey-Relevant-Excerpts.pdf>.

<sup>116</sup> See Kristen J. Osenga, The Loss of Injunctions Under eBay: Evidence of the Negative Impact on the Innovation Economy, Hudson Institute (Feb. 28, 2024), available at <https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impact-innovation-economy>.

business discussions—can be at best described as consistent with the “efficient infringement” approach employed by many Big Tech companies.

681. As explained earlier, “efficient infringement” (also known as “predatory infringement”) refers to the practice where large companies deliberately choose to infringe a patent rather than pay for a patent license, believing that the costs and hurdles stacked against a patent owner will deter it from enforcing its patent rights and believing that they are better off paying damages for past infringement in the form of a court-ordered compulsory license after years of delay resulting from the litigation process, rather than engaging in licensing negotiations prior to infringement for a fair, market-based fee.

682. Here NVIDIA extensively used Xockets’ patented technology, in deliberate disdain of Xockets’ legitimate patent rights, for which Xockets has been granted exclusive rights under United States patent laws. NVIDIA then refused to engage in negotiations with Xockets, yet has continued to reap profits from its infringing sales of Accused Products in a “trillion dollar” industry, dwarfing any court-ordered reasonable royalty that Xockets could possibly recover after years of protracted litigation. It is impossible to calculate the difference in the value Xockets would receive from an exclusive, negotiated license to its patented technology as compared to the amounts Xockets would collect in compulsory license payments through litigation if no injunctive relief is granted. This is the epitome of irreparable harm.

683. There is nothing efficient about the wider impact of “efficient infringement” on the patent system and innovation. “Efficient infringement” circumvents the fundamental constitutional promise of rewarding inventors for their inventive efforts to instead reward large companies for free-riding on the work of others. “[T]he loss of injunction to stop violations of property rights also devalues property in the marketplace—it is simply worth less given that it

offers less protection to its owner.”<sup>117</sup> Injunctive relief “serves both as a deterrent to patent infringement and facilitates market transactions in which fair market value is set through commercial negotiations.” *Id.*

684. As others have explained, efficient infringement “undermines the proper functioning of the patent system. It frustrates the promise of the reward to the innovator for one’s inventive labors. Once inventors know that the deck of (legal) cards is stacked against them and that they will suffer efficient infringement, they will create less patentable innovation. Without legal security in stable and effective property rights, venture capitalists will not invest in inventors or startups and the innovation economy will suffer.”<sup>118</sup>

685. “Efficient infringement is the cause of much distress and agony for innovators struggling to survive. The very existence of widespread efficient infringement, which is nothing more than stealing, absolutely stifles innovation.”<sup>119</sup>

686. Inventors who seek to enforce their patent rights against infringers through litigation face of ever-heightening hurdles: the high costs of patent litigation necessary to combat a “wars of attrition” that large corporations with comparatively vast resources will wage against them; duplicative and wasteful challenges to patentability before the Patent Office that turn the notion of double jeopardy on its head; and legal developments that continue to restrict the infringement remedies available to inventors.

---

<sup>117</sup> See Kristen J. Osenga, The Loss of Injunctions Under eBay: Evidence of the Negative Impact on the Innovation Economy, Hudson Institute (Feb. 28, 2024), available at <https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impact-innovation-economy>.

<sup>118</sup> <https://cip2.gmu.edu/2017/05/11/explaining-efficient-infringement>.

<sup>119</sup> <https://ipwatchdog.com/2017/03/17/tech-ruling-class-stifles-innovation-efficient-infringement/id=79391>.

687. The public has a critical interest in protecting the patent rights of innovators against the significant and growing threat of “efficient infringement.” Without stable and effective patent rights, investors will not back inventors, and our nation’s innovation economy will continue to decline in the face of competition from China and elsewhere. As retired Chief Judge Michel of the U.S. Court of Appeals for the Federal Circuit explained, “powerful tech companies have long relied on a strategy of deliberate infringement because enforcement litigation is too expensive for younger smaller competitors. . . . It’s no wonder the startup formation rate is at its lowest ebb in four decades. According to several economists, patent values, on average, fell 60 percent. Studies also found a sharp decline in venture-capital money going into real technology or deep tech changing our physical world, like computer chips, and shifting instead to entertainment and social media. Investment in entertainment and social media is faster, not as risky, and not dependent on patents.”<sup>120</sup>

688. Awarding injunctive relief also serves the public interest because “[a]bsent the threat of a permanent injunction, the incentives to ‘engage in the toils of scientific and technological research’, are reduced if not eliminated.” Acri née Lybecker, Kristina M.L., *Injunctive Relief in Patent Cases: the Impact of eBay* (June 14, 2024), available at <https://ssrn.com/abstract=4866108>; see also Kristen J. Osenga, *The Loss of Injunctions Under eBay: Evidence of the Negative Impact on the Innovation Economy*, Hudson Institute (Feb. 28, 2024), available at <https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impact-innovation-economy> (“The preliminary data demonstrates that patents are being devalued following *eBay*,” which many courts cite as supporting denials of injunctive relief).

---

120

[https://www.realclearpolicy.com/articles/2020/11/20/big\\_tech\\_is\\_overwhelming\\_our\\_political\\_system\\_650331.html](https://www.realclearpolicy.com/articles/2020/11/20/big_tech_is_overwhelming_our_political_system_650331.html).



Denying Xockets injunctive relief would effectively endorse devaluation of exclusive patent rights

[REDACTED]

[REDACTED] and then merely pay a compulsory license—effectively, a court-imposed non-exclusive license—when and if they are sued, and only after years of litigation.

689. Finally, enforcing inventors’ exclusive rights through the issuance of injunctive relief hews to the Founders’ vision for the patent system as reflected in the Constitution itself. The Founders recognized that enforcing inventors’ exclusive rights incentivizes innovation. As one scholar has aptly explained:

Property rights are essential to a free market, a growing innovation economy, and a flourishing society. The Founders recognized this basic truth and created the political and legal institutions necessary to restrain governmental power and protect the rights of life, liberty, and property. . . . Courts secure property rights, whether in land or in inventions, by doing more than ordering the payment of any damages caused by ongoing violations or willful infringement of the property. They also issue an injunction—an order backed by the coercive power of the state—that the defendant must stop committing the wrong. This remedy is necessary to secure the liberty interests of property owners in the free use of their property. It is also essential in ensuring that individuals will transact in the marketplace, as an injunction is the backstop to any negotiation. The power of a property owner to say “no” is the genesis of a negotiation in which individuals reach a meeting of the minds in exchanging goods and services at a freely negotiated market price.<sup>121</sup>

690. NVIDIA’s wrongful infringement has caused Xockets to suffer irreparable harm, resulting from the loss of its lawful patent rights to exclude others from making, using, selling, offering for sale, and importing the patented technology, as set forth in detail in the preceding paragraphs.

---

<sup>121</sup> <https://www.heritage.org/economic-and-property-rights/report/the-supreme-court-or-congress-must-restore-injunctions-patent>.

691. The balance of hardships clearly weighs in Xockets' favor. NVIDIA had notice of Xockets' Asserted Patents directly from inventor Parin Dalal on February 10, 2022. NVIDIA thus had an opportunity to negotiate for a license to the Asserted Patents, but instead chose to cease all further contact with Dr. Dalal while it continued and expanded the scope of its willful infringement.

692. Awarding injunctive relief serves the public interest because it holds infringers accountable and rewards innovation. Absent injunctive relief, infringers are incentivized to practice "efficient infringement" or "predatory infringement" as opposed to engaging in good faith licensing negotiations because failure to do so lacks any real consequence. This leads to weak and devalued patents, disincentivizes innovation and investment in innovation, and makes resolution of infringement disputes difficult—if not impossible—outside of costly and wasteful litigation.

### **DAMAGES**

693. Defendants' acts of infringement have caused damages to Xockets, and Xockets is entitled to recover from Defendants the damages sustained by Xockets as a result of Defendants' wrongful acts in an amount to be determined at trial.

694. Xockets is entitled to, and now seeks to, recover damages in an amount not less than the maximum amount permitted by law caused by Defendants' acts of infringement.

695. As a result of Defendants' acts of infringement, Xockets has suffered actual and consequential damages. To the fullest extent permitted by law, Xockets seeks recovery of damages in an amount to compensate for Defendants' infringement. Xockets further seeks any other damages to which Xockets would be entitled to in law or in equity.

### **ATTORNEYS FEES**

696. Xockets is entitled to recover reasonable and necessary attorneys' fees under applicable law.

**DEMAND FOR JURY TRIAL**

697. Pursuant to Rule 38 of the Federal Rules of Civil Procedure, Xockets demands a trial by jury on all issues so triable.

**PRAYER FOR RELIEF**

WHEREFORE, Xockets prays for judgment and requests that the Court find in its favor and against Defendants. Xockets respectfully requests that the Court enter preliminary and final orders, declarations, and judgments against Defendants as are necessary to provide Xockets with the following relief:

- a. A judgment that Defendants' conduct alleged above is in violation of Sections 1 and 2 of the Sherman Act;
- b. An award of threefold of the damages Plaintiff shall prove it has sustained on account of Defendants' violation of the antitrust laws, including, without limitation, pre-judgment and post-judgment interest;
- c. A judgment that Defendants have infringed and/or are infringing one or more claims of the Asserted Patents, literally or under the doctrine of equivalents, and directly or indirectly, as alleged above;
- d. A judgment that Defendants' infringement of the claims of the Asserted Patents has been willful;
- e. An award for all damages and costs arising out of Defendants' infringement, to adequately compensate Xockets for Defendants' infringement of the Asserted Patents, but in no event less than a reasonable royalty, including an accounting of damages up to any verdict as well as supplemental damages for any continuing post-verdict infringement up until entry of the final judgment, with an accounting, as needed;

- f. Pre-judgment and post-judgment interest, jointly and severally, in an amount according to proof;
- g. Treble damages based on Defendants' willful infringement;
- h. An accounting of damages and any future compensation due to Xockets for Defendants' infringement (past, present, or future) not specifically accounted for in a damages award (or other relief), and/or permanent injunctive relief;
- i. A judgment that this case is exceptional and an award of reasonable attorneys' fees as provided by 35 U.S.C. § 285 and enhanced damages as provided by 35 U.S.C. § 284;
- j. The entry of an order preliminarily enjoining and restraining Defendants and their parents, affiliates, subsidiaries, officers, agents, servants, employees, attorneys, successors, and assigns and all those persons in active concert or participation with them or any of them, from violating the antitrust laws and from making, importing, using, offering for sale, selling, or causing to be sold the Accused Products, including the Blackwell GPU-enabled server computer systems that fall within the scope of any claim of the Asserted Patents, or otherwise infringing or inducing infringement of any claim of the Asserted Patents;
- k. The entry of an order permanently enjoining and restraining Defendants and their parents, affiliates, subsidiaries, officers, agents, servants, employees, attorneys, successors, and assigns and all those persons in active concert or participation with them or any of them, from violating the antitrust laws and from making, importing, using, offering for sale, selling, or causing to be sold the Accused Products including the Blackwell GPU-enabled server computer systems that fall within the

scope of any claim of the Asserted Patents, or otherwise infringing or inducing infringement of any claim of the Asserted Patents; and

1. All further relief in law or in equity as the Court may deem just and proper.

Dated: September 5, 2024

Respectfully submitted,

/s/ Max Ciccarelli

**Max Ciccarelli** (SBN 00787242)

**CICCARELLI LAW FIRM**

100 N 6th Street, Suite 503

Waco, Texas 76701

Telephone: 214-444-8869

Email: max@ciccarellilawfirm.com

**Jason G. Sheasby** (*pro hac vice forthcoming*)

**IRELL & MANELLA LLP**

1800 Avenue of the Stars

Suite 900

Los Angeles, CA 90067

Tel.: 310.277.1010

Fax: 310.203.7199

Email: jsheasby@irell.com

**Jamie H. McDole** (SBN 24082049)

Lead Counsel

**Phillip B. Philbin** (SBN 15909020)

**Michael D. Karson** (SBN 24090198)

**David W. Higer** (SBN 24127850)

**Miranda Y. Jones** (SBN 24065519)

**Grant Tucker** (SBN 24121422)

**Matthew L. Vitale** (SBN 24137699)

**WINSTEAD PC**

2728 N. Harwood Street

Suite 500

Dallas, Texas 75201

Tel.: (214) 745-5400

Fax: (214) 745-5390

Email: jmcdoe@winstead.com

pphilbin@winstead.com

mkarson@winstead.com

dhiger@winstead.com

mjones@winstead.com

gtucker@winstead.com

mvitale@winstead.com

**Austin C. Teng** (SBN 24093247)

**Nadia E. Haghighatian** (SBN 24087652)

**WINSTEAD PC**

600 W. 5th Street

Suite 900

Austin, Texas 78701

Tel.: (512) 370-2800

Fax: (512) 370-2850

Email: ateng@winstead.com

nhaghighatian@winstead.com

**ATTORNEYS FOR PLAINTIFF**

**XOCKETS, INC.**

**CERTIFICATE OF SERVICE**

Defendant has not yet filed an appearance; Plaintiff will therefore serve a copy of this document on Defendants with the service of the Summons in this action.

/s/ Max Ciccarelli